AD_____

Award Number: DAMD17-94-J-4332

TITLE: Statistical Methods for Analyzing Time-Dependent Events in Breast Cancer Chemoprevention Studies

PRINCIPAL INVESTIGATOR: George Wong, Ph.D.

CONTRACTING ORGANIZATION: Strang Cancer Prevention Center
New York, New York 10021

REPORT DATE: October 1999

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release
distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

20001019 075

1

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 1999	Final (30 Sep 94 – 29 Sep 99)

4. TITLE AND SUBTITLE
Statistical Methods for Analyzing Time-Dependent Events in Breast Cancer Chemoprevention Studies

5. FUNDING NUMBERS
DAMD17-94-J-4332

6. AUTHOR(S)
George Wong, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Strang Cancer Prevention Center
New York, New York 10021

e-mail: gwong@strang.org

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release
distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)
The overall aim of our research proposal is the statistical non-parametric inferences of the redistribution-to-the-center estimator (RTCE) and the generalized maximum likelihood estimator (GMLE) for the survival function of a time-to-event variable that is subject to interval censoring. The RTCE, which is proposed by us, has a closed-form expression and is equal to GMLE under a homogeneous condition. The GMLE is the standard estimator in survival analysis. However, it cannot be expressed in a closed-form expression, and asymptotic distribution theory for it has been limited. From the study of the asymptotic properties of RTCE, we have gained important insight into proofs of asymptotic properties of GMLE. Specifically, we have established consistency, asymptotic normality and efficiency of GMLE under different conditions. Also, we have derived an asymptotic nonparametric two-sample distance test for comparing two populations. Under finite distributional assumptions on the survival and censoring distributions, we have established consistency, asymptotic normality for both the regression coefficients and the survival function of the Cox regression model. We point out major computational limitations associated with the Newton-Raphson algorithm for computing the asymptotic estimates of the Cox regression parameters, and suggest a simpler two-step estimation alternative.

14. SUBJECT TERMS
Breast Cancer, Interval-Censored Data, Consistency, Asymptotic Normality, Cox Regression

15. NUMBER OF PAGES
143

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

_____ Where copyrighted material is quoted, permission has been obtained to use such material.

_____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

_____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.
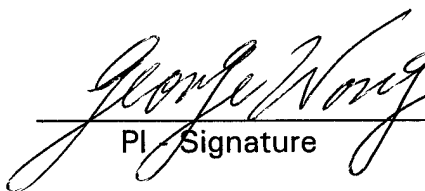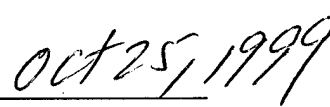
_____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

_____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

_____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

_____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

_____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

_____   _____
        PI / Signature   Date

A. TABLE OF CONTENTS

B. INTRODUCTION

In clinical follow-up studies, subjects are monitored at regular time intervals for a physical condition. It is often the case that an event under observation can take place in between two successive visits, and it may not be possible for the subject to know the time to such an event exactly. For example, consider the situation in which a group of women at high risk for breast cancer is asked to take a chemopreventive substance for a fixed time period. At the end of the period, each participating woman is required to submit a blood or urine sample at regular intervals in order to monitor the level of a validated intermediate biomarker. Let X denote the time from cessation of use of the agent to the loss of its protective effect, quantified as a return to baseline value of the biomarker. If a woman submits a sample for assay on a daily basis, the value of X can be observed exactly, unless the protective effect is still present by the time the study is terminated so that X is right censored in the usual sense of survival analysis. In practice, however, the follow-up interval can be a week or longer; therefore the exact value of X is generally unknown but is known to lie between the time points L and R, where L is the number of days from cessation of agent intake to the last time the sample was assayed and the protective effect was still present, and R is the number of days from cessation of agent intake to the most recent time the sample was assayed. If the protective effect is still present, then R takes the value infinity. In any case, when the value of X is only known to lie between (L, R), we say that X is censored in the interval (L, R). Therefore the observed data consist of either censoring intervals (L, R) or exact observations $X = L = R$.

Our research project is concerned with nonparametric estimation of the distribution function $F(t) = Pr(X \leq t)$ of a real-valued random variable X, or equivalently its survival function $S(t) = 1 - F(t)$, when the sample data are incomplete due to restricted observation brought about by interval censoring. Generalized maximum likelihood (GML) method in the sense of Kiefer and Wolfowitz [1] is the standard practice of estimating S. At present, there are two iterative computation procedures that will yield the GML estimate (GMLE) of S at convergence. The first one is due to Peto [2] and makes use of the Newton's method. The second is due to Turnbull [3] and makes use of a simpler but slower algorithm called self-consistent algorithm. A solution to this algorithm is also called a self-consistent estimator (SCE).

Because there is no closed-form expression for the GMLE of S, it has been difficult to study its asymptotic statistical properties, including consistency, normality and efficiency. Such a setback in the statistical development of the GMLE has severely limited its use in the statistical analysis of interval-censored (IC) data.

Before we began our funded Army research, we had extended Efron's redistribution-to-the-right idea for right-censored data [4] and proposed a redistribution-to-the-center (RTC) method to yield a nonparametric estimator of S which are called RTC estimate (RTCE).

Such an estimator has a closed-form expression and can be readily calculated for IC data of any dimension. IC data are said to satisfy DI (disjoint or included) condition if for every two censoring intervals, either they are disjoint or one is a subset of the other. For instance, in a clinical study in which every subject has the same follow-up schedule, say at time point a_1, a_2, ..., a_k, then $\{L, R\} = \{0, a_1\}$, or $\{a_i, a_{i+1}\}$ or $\{a_i, \infty\}$. A sample of such IC data $\{L_1, R_1\}, ..., \{L_n, R_n\}$ will satisfy DI condition. We had shown that under DI condition, RTCE is actually GMLE itself. This important observation, together with the availability of an explicit expression, had motivated us to submit the present proposal on RTCE to the Army.

In our first year of research, we completed our research for Task 1 and Task 2 in the Statement of Work for RTCE. However, we also discovered that in the case of non-DI data, RTCE may be different from GMLE, and RTCE is not always consistent. The interesting and intriguing observation is that the difference between RTCE and GMLE is small, at least based on our limited simulation studies [5]. In establishing consistency result for RTCE under DI condition, we had gained important insight into proofs of asymptotic properties for GMLE, which does not possess a closed-form expression. Because GMLE is the preferred estimator for S, we decided to focus our attention on GMLE instead of RTCE for the remainder of the funded research, and we have successfully completed all the tasks stated in the Statement of Work for GMLE.

Our research was then extended to study the statistical inferences with multivariate interval-censored data, which may also occurred in breast cancer research and Cox regression models. Some results have been obtained in these respects.

C. BODY

C.1. Basic setup

Interval-censored data can arise in the following four situations:

1. Case 2 IC data (C2 data) consist of right-censored ($R = \infty$), left-censored ($L = 0$) and strictly interval-censored observations ($0 < L < R < \infty$). These are by far the most common type of IC data in clinical follow-up studies.

2. Mixed IC data (MIC data) consist of both C2 data and exact observations ($L = R$). Yu, Li and Wong [6] presented an example involving MIC data from a breast cancer follow-up study.

3. Case 1 IC data (C1 data)) consist of either right-censored or left-censored observations. For example, when an animal is sacrificed for inspection of a tumor formation, time to appearance of the tumor is C1 interval censored. Examples of C1 data can be found in [7] and [8].

4. Doubly-censored data (DC data) consist of right-, left-censored and exact observations. An example with DC data is given in [9].

We have formulated four different interval censorship models corresponding to the four IC data types. To study the asymptotic properties of the GMLE, we make use of the following assumptions:

(AS1) The censoring distribution is discrete but the survival distribution is arbitrary.

(AS2) The support set of the censoring vector is finite, but the survival distribution is arbitrary.

(AS3) A probability restriction. See Section C.

(AS4) A probability restriction. See Section C.

(AS5) The censoring distribution and the survival distribution are arbitrary, but have to satisfy some regularity conditions, stated in Gu and Zhang [10].

C.2. Case 1 model

Case 1 model for C1 data assumes that the survival time X and a random inspection time Y are independent. We always observe Y. However, X is not fully observed except that we know that either $X \le Y$ or $X > Y$. Under assumption AS1, we have shown that GMLE is strongly consistent, asymptotically normal and asymptotically efficient at all the inspection times. The results are published in Yu, Schick, Li and Wong [11].

C.3. Case 2 model

The C2 model for C2 data assumes that X and the random censoring vector (Y, Z) are independent and that $Y < Z$ with probability one. We do not observe X except that we know X is before Y, or between Y and Z, or after Z. We state an assumption for C2 model as follows:

(AS3) $P\{X \in I_i \cap I_j\} > 0$ for any two realizations of (L, R), $(L_i, R_i) = I_i$ and $(L_j, R_j) = I_j$, provided $I_i \cap I_j \ne \emptyset$.

Under the assumption AS1, we have shown that GMLE is strongly consistent. Under the assumptions AS2 and AS3, we have shown that GMLE is asymptotically normal and efficient. The results are published in Yu, Schick, Li and Wong [12].

C.4. MIC model

Mixture interval censorship (MIC) model for MIC data assumes that an IC observation is drawn from a probability mixture of C2 model and the usual right censorship model for right-censored data.

Define $\tau = \sup\{t; \ Pr(\min(X, T) \le t) < 1\}$, $\tau_Y = \sup\{t; \ Pr(Y \le t) = 0\}$. and $\tau_Z = \sup\{t; \ Pr(Z \le t) < 1\}$. We assume that $\tau \ge \tau_Z$. We state an assumption for MIC model as follows:

(AS4) $\Pr(L = \tau) > 0$ if $Pr(X < \tau) < 1$ and $\Pr(R = \tau_Y) > 0$ if $Pr(X \le \tau_Y) > 0$.

Under assumptions AS2 and AS4, we have shown that GMLE is strongly consistent (Yu, Li and Wong [6]), and under assumptions AS2, AS3 and AS4, GMLE is asymptotically normal (Yu, Li and Wong [13]). Recently, we have been able to establish these asymptotic properties without the need of assumption AS2. A manuscript on these results has been

submitted for publication (Yu, Li and Wong [14]).

C.5. DC model

The DC model for DC data assumes that X and a random vector (Y, Z) are independent and $Y < Z$ with probability one, and that X is uncensored if $Y < X \leq Z$, right censored if $Z < X$ and left censored if $X \leq Y$. Let S_Z and S_Y be the survival functions of Z and Y, respectively, and let $K = S_Y - S_Z$. We state an assumption for DC model as follows:

(**AS5**) $\qquad K(x-) > 0$ for all x such that $S(x) < 1$ and $S(x-) > 0$,

We have shown in a submitted manuscript (Yu and Wong [15]) that in order to establish asymptotic results, GMLE has to be modified. Under assumptions AS4 and AS5 we have shown that the modified GMLE is strongly consistent and is asymptotically normal and efficient under assumptions AS3, AS4 and AS5.

C.6. Two-sample nonparametric test

Based on the asymptotic results that we have established for different IC models, we have successively derived the asymptotic distribution of the following two-sample distance test statistics for each model:

$$D = \int_{\tau_1}^{\tau_2} W(t)(\hat{S}_1(t) - \hat{S}_2(t))dt,$$

where τ_1 and τ_2 are specified time point and $W(t)$ is a weight function. A manuscript on the asymptotic results of D is being prepared.

C.7. Proportional hazards model

In our original proposal, we had assigned three months of time for Task 7 on Cox regression for IC data. However, we have realized that statistical inference for the parameter $\underline{\beta}$ in Cox regression under interval censorship is much more involved than its counterpart in the usual right-censored situation. In the latter case, the maximum likelihood estimator (MLE) of $\underline{\beta}$ does not depend on the baseline survival function $S_0(t)$ owing to the simple nature of the partial likelihood approach. However, such simplicity of likelihood function does not carry over to the interval censorship model, and maximum likelihood estimation of $\underline{\beta}$ will involve GML estimation of $S_0(t)$ at the same time, thus resulting in a difficult high-dimensional estimation problem.

Under the restrictive assumption that both X and the censoring vector take on finitely many values, we have proved that the MLE of $\underline{\beta}$ and the GMLE of $S_0(t)$, and hence the survival function $S(t|\underline{Z}) = S_0(t)exp^{\underline{\beta}'\underline{Z}}$, where $\underline{Z}$ denotes a vector of covariates for Cox regression, are consistent and asymptotically normal (Li, Yu and Wong [17]). Much more effort is needed to pursue research on the asymptotic inference of Cox regression model under more relaxed assumptions on the distributions of X and the censoring vector.

During the no-cost extension period, we have devoted our effort to the implementation of a Newton-Raphson algorithm for computing the MLE of $\underline{\beta}$ and the GMLE of $S_o(t)$. Although the algorithm is straightforward to derive using the asymptotic covariance matrix which we have derived for the Cox parameters, we soon realized there are two difficult problems associated with the Newton-Raphson algorithm. The first problem is that the algorithm is computationally infeasible for data of moderate size. For example, in the prognostic analysis of a breast cancer relapse follow-up study with $n = 374$ women which we shall describe in Section C.9, the Newton-Raphson algorithm broke down owing to the numerical difficulty associate with inverting a Hessian matrix of order 60. Another problem with the Newton-Raphson algorithm is that it does not guarantee the strict monotonicity condition $S_o(t_1) > \cdots > S_o(t_m)$ is satisfied at each iteration, where t_1, ..., t_m are the ordered distinct times points. When this condition is violated, we shall have to re-compute the estimates by assuming $S_o(t_j) = S_o(t_k)$ for some $j \neq k$. Since there are a maximum of 2^m such possibilities, it will be computationally infeasible to apply the Newton-Raphson algorithm to a data set with even a moderate m.

The above computational problems associated with the Newton-Raphson algorithm have motivated us to consider a two-step estimation approach for the Cox regression parameters. Briefly, in step 1, the regression coefficient are estimated by a simple Newton-Raphson algorithm through the device of a data grouping scheme; in step 2, the baseline survival function is estimated by a simple self-consistent algorithm based on the original data. The details of our novel approach are contained in the DOD grant "Cox regression model for interval-censored data in breast cancer follow-up studies", which we have submitted to the USAMRMC for consideration for funding.

C.8. Computer software

We have made it available to the public a set of computer programs for calculating RTCE and GMLE, for carrying out asymptotic inference of GMLE for all patterns of interval censorship, and for evaluating the Z-score of the proposed two-sample weighted distance test. These programs can be accessed via the internet at qyu@math.binghamton.edu.

C.9. Applications to breast cancer research

We have applied our results on asymptotic inference of GMLE for C2 model to two breast cancer research projects. The first project is concerned with a chemoprevention intervention trial of indole-3-carbinol (I3C) for breast cancer which is being conducted at Strang Cancer Prevention Center. The statistical question of interest is the estimation of duration of sustaining effect of I3C, which is C2 censored. A preliminary report on a short-term trial has recently been published [18]; however, a longer trial lasting for more than one year is still ongoing so that more informative data on duration of sustaining effect can be obtained.

The second project is a breast cancer relapse follow-up study based on data obtained

9

from 374 women with stages I - III unilateral invasive breast cancer surgically treated at Memorial Sloan-Kettering Cancer Center between 1985 and 1990. The median follow-up duration was 46 months. Relapse time was given by the time interval between surgery and the initial relapse. A relapse that took place between two successive follow-up visits was regarded as interval censored. If a patient did not relapse towards the end of the study, then her relapse time was right censored. Of the 374 observations, 300 were right censored (no relapse), 21 were left censored and 53 were strictly interval censored (74 relapses). Bone marrow micrometastasis (BMM) was determined for each woman at the time of surgery. An important question is whether remission duration is related to the extent of initial tumor burden defined as number of BMM cells detected. Figure 1 compares the relapse-free GMLE curves of patients with number of BMM ≤ 14 versus those with number of BMM > 14. Our asymptotic two-sample distance test yielded a P value close to 0.1. An abstract on a detailed prognostic analysis of the entire data set using our asymptotic results on C2 data was presented at the annual San Antonio Breast Cancer Symposium in December 1998.

D. KEY RESEARCH ACCOMPLISHMENTS

- presented a simple nonparametric estimator of the survival function called RTCE, which has an explicit expression and which is equal to GMLE under some restrictions on the interval-censored data

- established consistency, asymptotic normality and asymptotic efficiency of GMLE under a variety of interval censorship models

- presented an asymptotic two-sample nonparametric test for different interval censorship models

- established consistency, asymptotic normality and asymptotic efficiency for the MLE of the regression coefficients and GMLE of the survival function at a given covariate pattern of a Cox regression model under finite assumptions on the distribution functions of both the survival time and the censoring vector

- identified the computational difficulties associated with the Newton-Raphson algorithm for computing the asymptotic estimates of Cox parameters

- pointed out future directions for a more feasible asymptotic Cox regression analysis of interval-censored data

- made available to the public a set of computer programs for calculating RTCE and GMLE, carrying out asymptotic inference of GMLE and for evaluating the Z-score of the proposed two-sample nonparametric test

- applied the established asymptotic generalized maximum likelihood results successfully to a breast cancer relapse follow-up study with 374 women

E. REPORTABLE OUTCOMES

- **10 published articles:**

[a] Li, L., Watkins, T. and Yu, Q. (1997). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. *Scandinavian Journal of Statistics.* 24, 531-542.

[b] Yu, Q., Li, L. and Wong, G. Y. C. (1999). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scandinavian Journal of Statistics.* (In press).

[c] Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics.* Vol. 4.

[d] Yu, Q., Li, L. and Wong, G. Y. C. (1998). Asymptotic variance of the GMLE of a survival function with interval-censored data. *Sankhya,* A. 60, 184-197.

[e] Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Statistics & Probability Letters* 37, 223-228.

[f] Yu, Q. and Wong, G. Y. C. (1998). Consistency of self-consistent estimators of a discrete distribution function with bivariate right-censored data. *Communication in Statistics.* 27, 1461-1476.

[g] Wong, G. Y. C. and Yu, Q. (1999). Generalized MLE Of a joint distribution function with multivariate interval-censored data. *Journal of Multivariate Analysis* 69, 155-166.

[h] Schick, A. and Yu, Q. (1999). Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics.* (In press).

[i] Li, L. and Yu, Q. (1997). Self-consistent estimators of survival functions with doubly-censored data. *Communication in Statistics,* 2609-2623.

[j] Wong, G. Y. C., Bradlow, H. L., Sepkovic, D., Mehl, S., Mailman, J. and Osborne, M. P. (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplements* 28/29, 111-116.

Copies of the articles are included in APPENDICES.

- **2 submitted manuscripts:**

[a] Yu, Q., Li, L. and Wong, G. Y. C. Asymptotic properties of NPMLE with mixed interval-censored data. (Submitted to the *Annals of the Institute of Statistical Mathematics*)

[b] Yu, Q. and Wong, G. Y. C. A modified GMLE with doubly-censored data. (Submitted to *Australian Journal of Statistics*).

- **7 abstract presentations:**

[a] Q. Yu, G, Y.C. Wong and L. Ye. Estimation of a survival function with interval-censored data, a simulation study on the redistribution-to-the-inside estimator. *1995 Joint sta-*

tistical meetings at Orlando, Florida, U.S.. August 13-17, 1995.

[b] Q. Yu, L. Li and G.Y.C. Wong (1996) Variance of the MLE of a survival function with interval-censored data. *1996 Sydney international statistical congress, Australia.* July 8-12, 1996.

[c] Q. Yu, L. Li and G.Y.C. Wong (1996) Variance of the MLE of a survival function with doubly-censored data. *1996 Joint statistical meetings at Chicago, Illinois, U.S..* August 4-8, 1996.

[d] Q. Yu and L. Li. Asymptotic properties of self-consistent estimators with doubly-censored data. *1997 Joint statistical meetings at Anaheim, California, U.S..* August 10-14.

[e] Yu, Q. and G.Y.C. Wong. Asymptotic Properties Of Self-Consistent Estimators of A Survival Function *ICSA 1997 Applied Statistics Symposium at Rutgers University, New Jersey, U.S..* May 30 - June 1, 1997.

Copies of the abstracts are included in APPENDICES.

- **computer programs** for asymptotic inferences of GMLE at the internet site QYU@math.binghamton.edu
- **a proposal** entitled "Statistical analysis of multivariate interval-censored data in breast cancer follow-up studies" based on work support by this award has been funded by USAMEMC from 7/1/99 to 6/30/02 to George Y. C. Wong as principal investigator.
- **a proposal** entitled "Cox regression model for interval-censored data in breast cancer follow-up studies" based on work supported by this award has been submitted to USAMRMC since June 15, 1999 with George Y.C. Wong as the principal investigator, and Qiqing Yu as co-investigator.

F. CONCLUSIONS

In the four years of our DOD grant, we have successfully accomplished our research objectives on the asymptotic inference of the GMLE of the survival function for interval-censored data. Under different interval censorship models, we have established consistency, asymptotic normality and asymptotic efficiency of the GMLE. When both the survival time and the censoring vector take on finitely many values, we have established similar asymptotic properties for the maximum likelihood estimators of the regression coefficients and the GMLE of the survival function at a given covariate pattern of the Cox regression model for interval-censored data. We have made available to the public a set of computer programs for carrying out the asymptotic generalized maximum likelihood inference procedures for all types of interval-censored data. The results from our research will provide clinicians and basic science researchers in breast cancer with a set of fundamentally important statistical tools for the analysis of interval-censored data that are encountered in breast cancer chemoprevention studies, and relapse follow-up studies in which the time-to-event variable cannot be exactly observed.

12

Our research also indicates that asymptotic inferences for the parameters of the Cox regression model for interval-censored data cannot be feasibly obtained by a standard iterative algorithm, such as the Newton- Raphson algorithm. Our investigations into Cox regression in this grant have inspired us to consider a computational simpler two-step estimation procedure for the parameters of the Cox model. We have consolidated our ideas into a proposal entitled "Cox regression model for interval-censored data in breast cancer follow-up studies", which has been submitted to the USAMRMC for consideration for funding.

G. REFERENCES

[1] Kiefer, J and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* 27, 887-906.

[2] Peto, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.* 22, 86-91.

[3] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B, 38, 290-295.

[4] Efron, B (1967). The two sample problem with censored data. *Fifth Berkeley Symposium on Mathematical Statistics.* University of California Press, 831-853.

[5] Wong, G. Y. C. and Yu, Q. (1995). Estimation of a survival function with interval-censored data; A simulation study. (Preprint).

[6] Yu, Q., Li, L. and Wong, G. Y. C. (1997). On consistency of the self-consistent estimator of survival functions with interval censored data. *Scan. J. of Statist.* (Accepted).

[7] Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.* 26, 641-647.

[8] Keiding. N. (1991) Age-specific incidence and prevalence: A statistical perspective (with discussion) *JRSS,* A. 154, 371-412.

[9] Leiderman, P.H., Babu, D., Kagia, J., Kraemer, H.C. and Leiderman, G.F. (1973). African infant precocity and some social influences during the first year. *Nature,* 242, 247-249.

[10] Chang, M.N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* 15, 1536-1547.

[11] Yu, Q., Schick, A., Li, L. and Wong, G. Y.C. (1998). Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. *Canadian Journal of Statistics.* Vol. 4.

[12] Yu, Q., Schick, A., Li, L. and Wong, G. Y. C. (1998). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Prob. & Statist. Letters.* 37, 223-228.

[13] Yu, Q., Li, L. and Wong, G. Y. C. (1998). Asymptotic variance of the GMLE of a survival function with interval-censored data. *Sankhya,* A. 60. 184-197.

[14] Yu, Q., Li, L. and Wong, G. Y. C. Asymptotic properties of NPMLE with mixed interval-censored data. Submitted to *Annals of the Institute of Statistical Mathematics*).

[15] Yu, Q., and Wong, G. Y.C. A modified GMLE with doubly-censored data. (Submitted to *Australian J. of Statist.*).

[16] Li, L., Yu, Q. and Wong, G. Y. C. (1998). Proportional hazards model with interval-censored and exact observations. (Under preparation).

[17] Wong, G. Y. C., Bradlow, H. L., Sepkovic, D., Mehl, S., Mailman, J. and Osborne, M. P. (1997). A dose-ranging study of indole-3-carbinol for breast cancer prevention. *Journal of Cellular Biochemistry Supplements* 28/29, 111-116.

H. APPENDICES

a.   Figure 1

                page   16
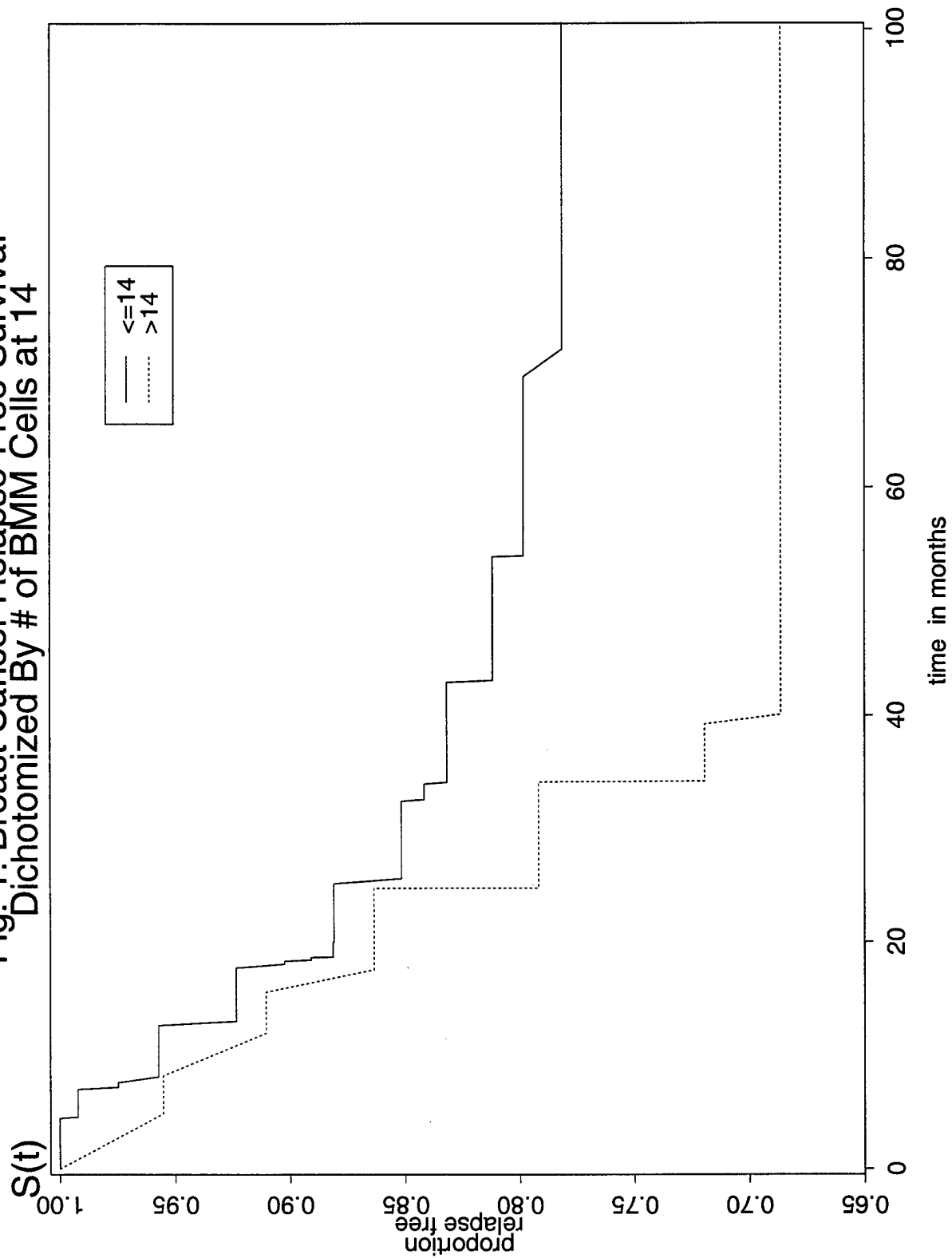
b.   10 published articles

                pages   17 – 107

c.   9 published abstracts

                pages   108 – 121

d.   2 submitted manuscripts

                pages   122 – 143

Fig. 1. Breast Cancer Relapse-Free Survival
Dichotomized By # of BMM Cells at 14

An EM Algorithm for Smoothing the Self-consistent Estimator of Survival Functions with Interval-censored Data

LINXIONG LI, TERRY WATKINS

University of New Orleans

QIQING YU

State University of New York at Binghamton

ABSTRACT. Interval-censored data arise in a wide variety of application and research areas such as, for example, AIDS studies (Kim *et al.*, 1993) and cancer research (Finkelstein, 1986; Becker & Melbye, 1991). Peto (1973) proposed a Newton–Raphson algorithm for obtaining a generalized maximum likelihood estimate (GMLE) of the survival function with interval-censored observations. Turnbull (1976) proposed a self-consistent algorithm for interval-censored data and obtained the same GMLE. Groeneboom & Wellner (1992) used the convex minorant algorithm for constructing an estimator of the survival function with "case 2" interval-censored data. However, as is known, the GMLE is not uniquely defined on the interval $[0, \infty)$. In addition, Turnbull's algorithm leads to a self-consistent equation which is not in the form of an integral equation. Large sample properties of the GMLE have not been previously examined because of, we believe, among other things, the lack of such an integral equation. In this paper, we present an EM algorithm for constructing a GMLE on $[0, \infty)$. The GMLE is expressed as a solution of an integral equation. More recently, with the help of this integral equation, Yu *et al.* (1997a, b) have shown that the GMLE is consistent and asymptotically normally distributed. An application of the proposed GMLE is presented.

Key words: generalized maximum likelihood estimator, EM algorithm, interval censorship, self-consistency

1. Introduction

Interval-censored data are frequently seen in medical studies, pharmaceutical applications, and engineering research. Let $X_1, X_2, \ldots, X_n$ denote a random sample of observations of a random variable X, called the failure time, with distribution function F, and let $(Y_1, Z_1), (Y_2, Z_2), \ldots, (Y_n, Z_n)$ denote a random sample of observations of a random vector (L, R), called the censoring vector, with joint distribution function $G(l, r)$, where with probability one, $L \leqslant R$. As is common, define $S(x) = 1 - F(x)$ as the survival function of F. For each observation X_i, there is a corresponding censoring vector (Y_i, Z_i). The failure time X_i is observed if it is outside the open interval (Y_i, Z_i). When X_i is within (Y_i, Z_i), we only observe (Y_i, Z_i) but not the value X_i, i.e. X_i is censored. When Z_i (Y_i) equals ∞ (0), the failure time X_i is subject to a right (left) censorship. If only $\min\{\max\{X_i, Y_i\}, Z_i\}$ is observed, we say the failure time is subject to a double censorship. It is readily seen that the interval censoring scheme contains right censoring and left censoring schemes as special cases. If the functional form of the distribution function F is known, we only need to estimate the parameters of F. However, when the functional form of F is unknown, a non-parametric approach must be used. This paper focuses on the latter.

Kaplan & Meier (1958) proposed the product limit estimator (PLE) to estimate the survival function when data are right-censored. There have been extensive studies concerning the PLE.

Doubly-censored data, which treat right-censored and left-censored data as special cases, are investigated by many authors. A self-consistent estimator (Efron, 1967) of the survival function with doubly-censored data as well as various properties of the estimator such as strong convergence, asymptotic normality, etc., are established (see, for example, Turnbull, 1974; Chang, 1990; Gu & Zhang, 1993). The self-consistent estimator is implicitly expressed as a solution of an integral equation. No closed forms of the estimator have been presented. For arbitrarily interval-censored data, Peto (1973) proposes a Newton–Raphson algorithm to obtain a generalized maximum likelihood estimator (GMLE) (see Kiefer & Wolfowitz, 1956; Johansen, 1978) of the survival function. Turnbull (1976) derives a self-consistent algorithm and shows that the algorithm converges monotonically to the GMLE. This GMLE is, however, not uniquely determined in innermost intervals (see definition below). Furthermore, Turnbull's self-consistent equation is not in the form of an integral equation. Studies about arbitrarily interval-censored data are not as fruitful as those mentioned above due to, among other things, lack of an integral equation for the GMLE. Tsai & Crowley (1985) discuss connections among the GMLE, the EM algorithm, and the self-consistent estimators for incomplete data, focusing on right censoring and double censoring cases, taking advantage of availability of the integral equations for the latter two models. Groeneboom & Wellner (1992) use the convex minorant algorithm for computing the MLE of the survival function with "case 2" interval-censored data. The "case 2" interval censoring is the same as arbitrary interval censoring described above except that the exact observations can never be observed (thus it is a special case of the arbitrary interval censoring). Yu & Wong (1996a, b) consider a special case of interval-censored data. They assume that any two censoring intervals are either disjoint or one includes another. This assumption covers a wide variety of situations. They derive an explicit expression for the GMLE of the survival function and then prove that the estimator is strongly consistent.

Since Turnbull's self-consistent GMLE is not uniquely defined on innermost intervals, it is not convenient to use the estimator if the data are heavily censored. In this paper, we propose an EM approach to construct a GMLE that is defined on the interval $[0, \infty)$. This approach also gives an integral equation expression for the GMLE. More recently, with the help of this integral equation, Yu et al. (1997a, b) prove the uniform strong consistency and asymptotic normality of the GMLE.

The organization of the paper is as follows. Section 2 provides the necessary definitions and background. In section 3, we prove the convergence of the proposed EM algorithm and show that it converges to the same GMLE as Turnbull's. An application of the estimator derived is presented in section 4.

2. Algorithms

Following the notation of section 1, assume that the vectors (Y_i, Z_i, X_i), $i = 1, \ldots, n$, are mutually independent, and that X_i and (Y_i, Z_i) are also independent. If $Y_i < X_i < Z_i$, the censoring interval (Y_i, Z_i) rather than the failure time X_i is observed and we denote the observation by an open interval $(L_i, R_i) = (Y_i, Z_i)$; if X_i is outside (Y_i, Z_i) we observe the exact failure time. In the latter case, we define $L_i = X_i = R_i$, and call X_i or the closed interval $[L_i, R_i]$ an exact observation. Thus we may assume that the final observations are

$$\{L_i, R_i\} = \begin{cases} [L_i, R_i] & \text{if } L_i = R_i \\ (L_i, R_i) & \text{if } L_i < R_i \end{cases}$$

(note: some of the intervals are collapsed to points), and, without loss of generality (WLOG), assume that $L_1 \leqslant L_2 \leqslant \cdots \leqslant L_n$. Let $\mathscr{L}$ denote the set $\{l_i, 1 \leqslant i \leqslant n\}$ and $\mathscr{R}$ the set

$\{r_i, 1 \le i \le n\}$, where $\{l_i, r_i\}$ are the realizations of $\{L_i, R_i\}$. Ranking the $2n$ points (n ls and n rs) in increasing order yields a sequence, say $c_1 \le c_2 \le \cdots \le c_{2n}$. If there exist ties in the observations, we suppose that

1. R_i has smaller rank than L_j if $R_i = L_j < R_j$;
2. L_i has larger rank than R_j if $L_i = R_j > L_j$;
3. If $\{L_i, R_i\} = \{L_j, R_j\}$ and $i < j$ then they are ranked as $L_i \le L_j \le R_j \le R_i$.

Define an innermost interval $\{p, q\}$ to be the non-empty intersection of observed intervals $\{L_i, R_i\}$ such that $\{p, q\} \cap \{L_i, R_i\}$ is either an empty set or $\{p, q\}$. Notice that every exact observation comprises a closed innermost interval and that distinct innermost intervals are disjoint. Suppose that there are m ($\le n$) such distinct (open or closed) innermost intervals: $\{p_1, q_1\}, \ldots, \{p_m, q_m\}$, where $p_1 \le q_1 \le \cdots \le p_m \le q_m$ and

$$\{p_i, q_i\} = \begin{cases} (p_i, q_i) & \text{if } p_i < q_i \\ [p_i, q_i] & \text{if } p_i = q_i. \end{cases}$$

Turnbull (1976) provides a self-consistent algorithm for obtaining the GMLE of S, and shows that the GMLE assigns weight on innermost intervals only. Specifically, define an indicator function $\delta_{ij} = 1$ if $\{p_j, q_j\} \subset \{l_i, r_i\}$, and 0 otherwise. Let

$$\mu_{ij}(s) = \frac{\delta_{ij} s_j}{\sum_{k=1}^{m} \delta_{ik} s_k}$$

where $s = (s_1, \ldots, s_m)$ are the masses assigned to the corresponding innermost intervals, satisfying $\sum_{i=1}^{m} s_i = 1$, $s_i \ge 0$, $1 \le i \le m$. Write

$$\pi_j = \frac{1}{n} \sum_{i=1}^{n} \mu_{ij}(s).$$

The GMLE, and hence the self-consistent estimator of S, can be obtained by the following iterative procedure.

1. Set the initial values $s_j^0 = 1/m$, $1 \le j \le m$.
2. Compute $\mu_{ij}(s)$, and set $s_j^1 = \pi_j(s^0)$.
3. Repeat step 2 by replacing s^0 with s^1, and so on.

This procedure converges monotonically to the estimate of the weight s. Although the GMLE of $S(x)$ can be formed as

$$\hat{S}(x) = \sum_{i, q_i > x} s_i, \quad x \notin \cup_j \{p_j, q_j\} \tag{2.1}$$

we only know the amount of weight on innermost intervals but not the way that the weight varies within the innermost intervals. We now present an EM algorithm for obtaining the GMLE of $S(x)$. The proposed GMLE assigns the same weight on innermost intervals as Turnbull's and describes the distribution of the weight within the innermost intervals. Meanwhile, an expression for the GMLE is obtained.

Let $H_0(x)$ denote a strictly increasing initial distribution function on $[0, a)$, where $a \ge \max\{r_i; r_i \in \mathscr{R}\}$, (for example, $H_0(x) = 1 - \exp(-x)$, $x \ge 0$. A choice of the initial H_0 is given in section 4) and define

$$H_1(x) = \frac{1}{n} \left[\sum_{i=1}^{n} \frac{H_0(x) - H_0(l_i)}{H_0(r_i-) - H_0(l_i)} I(x \in (l_i, r_i)) + \sum_{i=1}^{n} I(x \in [r_i, \infty)) \right]$$

where $I(A)$ denotes the indicator function of the event A and $f(x-) = \lim_{t \uparrow x} f(t)$. Then define a distribution function

$$H_2^i(x) = \frac{H_1(x) - H_1(l_i)}{H_1(r_i-) - H_1(l_i)} I(x \in (l_i, r_i)) + I(x \in [r_i, \infty)).$$

In other words, we truncate distribution H_1 on each censored interval. Let

$$H_2(x) = \frac{1}{n} \sum_{i=1}^{n} H_2^i(x).$$

Then use H_2 as an initial distribution function and repeat the above procedure to obtain H_3. More specifically, on the kth iterative procedure, H_k is calculated by

$$H_k(x) = \frac{1}{n} \sum_{i=1}^{n} H_k^i(x),$$

where, when $l_i = r_i$, $H_k^i(x) = 0$ if $x < l_i$, 1 if $x \geq l_i$, and, when $l_i < r_i$,

$$H_k^i(x) = \frac{H_{k-1}(x) - H_{k-1}(l_i)}{H_{k-1}(r_i-) - H_{k-1}(l_i)} I(x \in (l_i, r_i)) + I(x \in [r_i, \infty)).$$

In terms of conditional expectation,

$$H_k(x) = E_{H_{k-1}} \left[\frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x) | \{L_i, R_i\}, i = 1, \ldots, n \right].$$

This is an EM algorithm (Tsai & Crowley, 1985). A proof of the convergence of the EM algorithm is given in the next section. Thus, the limiting distribution, say H, is a self-consistent estimator of F. It is known that Turnbull's algorithm is also an EM algorithm. The difference of these two EM algorithms is previously described in the paragraph following the definition of Turnbull's algorithm. In addition, it is easy to see that in terms of convergence rate one is not superior to the other.

3. Main results

We first make sure that the proposed EM algorithm is well-defined, namely we need to guarantee that the denominators involved in H_k are not zero. This is assured by the following lemma. The proof of it is simply by induction on k and is omitted.

Lemma 1
For $k \geq 1$, $H_k(r_i-) - H_k(l_i) \geq 1/n$.

We now prove that the EM algorithm converges. The proof is similar to th. 2.1 of Tsai & Crowley (1985).

Theorem 1
As $k \to \infty$, $H_k(x)$ converges to, say $H(x)$.

Proof. By the definition of the EM algorithm, the initial estimator H_1 has its weight on observations $\{L_i, R_i, i = 1, \ldots, n\}$ only. This implies that it is the EM algorithm for incomplete multinomial data, which belong to exponential family. Thus by th. 2 of Wu (1983) the EM algorithm converges.

We now consider a transformation of the observed censoring intervals. The transformed data make the proofs simpler and produce the same self-consistent estimate as do the original data. Let $\{l_i, r_i\}$, $1 \leq i \leq n$, be the original data, and let $\{p_i, q_i\}$, $1 \leq i \leq m$, be the innermost intervals. For convenience, define $p_{m+1} = \infty$ and $q_0 = 0$. The transformation proceeds as follows. For any i, (1) if there is not any exact observation at q_i, then move all rs between q_i and p_{i+1} to q_i, (2) otherwise, move all rs between q_i and p_{i+1} to the smallest r that is greater than q_i; similarly, if there is not any exact observation at p_i, then move all ls between p_i and q_{i-1} to p_i, otherwise, move all ls between p_i and q_{i-1} to the largest l that is smaller than p_i. We call the transformation S-transformation. We use $\{L'_i, R'_i\}$ to denote the S-transformed data. To illustrate the transformation, consider the following example.

Example 1. If the original data are

$$(_1 \ (_2 \ (_3 \ x_4 \ )_2 \ (_5 \ )_5 \ )_1 \ x_6 \ (_7 \ )_3 \ x_8 \ )_7$$

where $(_j \ )_j$ denotes the jth censoring interval, then the S-transformed data are

$$(_{1,2,3} \ x_4 \ )'_2 \ (_5 \ )'_{5,1} \ x_6 \ (_7 \ )'_3 \ x_8 \ )'_7$$

or $l'_1 = l'_2 = l'_3 = l_3 < x_4 < r'_2 = r_2 \leq l'_5 = l_5 < r'_5 = r'_1 = r_5 < x_6 < l'_7 = l_7 < r'_3 = r_3 < x_8 < r'_7 = r_7$ (we pretend that $l'_1 \leq l'_2 \leq l'_3$ and $r'_5 \leq r'_1$).

It is important to note that the S-transformation does not change the innermost intervals and (L, R) contains an innermost interval if and only if (L', R') contains the same innermost interval. Noting that the likelihood function can be written as

$$\pounds = \prod_{i=1}^{n} \left(\sum_{k=1}^{m} \delta_{ik} s_k \right) \quad \text{(see Peto, 1973),}$$

we see that the S-transformation does not change the likelihood function. Since the GMLE of **s** is uniquely determined by, and has weight only on, innermost intervals (Peto, 1973), the original data and the corresponding S-transformed data produce the same GMLE of **s** (Yu & Wong, 1996a). Hence, from now on, we use the following convention.

Convention
We suppress the word S-transformation and assume that the data are already S-transformed unless otherwise specified.

Notice that the GMLE of S is entirely determined by **s** (see (2.1)), and thus the original data and the transformed data give the same GMLE of S.

Theorem 2
Suppose that the initial distribution H_0 is strictly increasing on $[0, a)$ where a is defined in section 2. Let $\mathcal{K} = \bigcup_{i=1}^{m} \{p_i, q_i\}$ and let $\mathscr{C}$ denote the support of the limiting distribution function $H = \lim_{k \to \infty} H_k$. Then $\mathscr{C} \subset \mathcal{K}$.

Proof. It is sufficient to prove that non-innermost intervals do not have weight. If (c_m, c_{m+1}) is a non innermost interval, then it must be one of the following cases:

(a) (l_j, x_{j+1}), where $l_j < x_{j+1} < r_j$; (remember the convention, i.e. there is no additional l_i or r_i within $(l_j, x_{j+1}]$)

(b) (x_j, l_{j+1});

(c) (x_j, r_i), where $i < j$ and $l_i < x_j < r_i$;

(d) (r_i, x_j);

(e) (r_p, l_j), where $l_p < r_p < l_j < r_j$.

We now prove that none of the above non-innermost intervals has weight. First consider (a). Note that

$$H(x_{j+1}-) - H(l_j) = \frac{1}{n} \sum_{i=1}^{j} \frac{H(x_{j+1}-) - H(l_j)}{H(r_i-) - H(l_i)} I(l_j \in (l_i, r_i))$$

$$= [H(x_{j+1}-) - H(l_j)] \frac{1}{n} \sum_{i=1}^{j} \frac{I(l_j \in (l_i, r_i))}{H(r_i-) - H(l_i)}$$

Thus, either $H(x_{j+1}-) - H(l_j) = 0$, or

$$H(x_{j+1}-) - H(l_j) \neq 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{j} \frac{I(l_j \in (l_i, r_i))}{H(r_i-) - H(l_i)} = 1. \qquad (3.1)$$

In addition, if (3.1) is true,

$$H(x_{j+1}) - H(l_j) = [H(x_{j+1}) - H(l_j)] \frac{1}{n} \sum_{i=1}^{j} \frac{I(l_j \in (l_i, r_i))}{H(r_i-) - H(l_i)}$$

$$+ \frac{1}{n} \sum_{i=1}^{n} [I(x_{j+1} \in [r_i, \infty)) - I(l_j \in [r_i, \infty))]$$

$$\geqslant [H(x_{j+1}) - H(l_j)] \frac{1}{n} \sum_{i=1}^{j} \frac{I(l_j \in (l_i, r_i))}{H(r_i-) - H(l_i)} + \frac{1}{n}$$

$$= [H(x_{j+1}) - H(l_j)] + \frac{1}{n}$$

which is impossible. Thus, $H(x_{j+1}-) - H(l_j) = 0$.

Now consider (b) $(c_m, c_{m+1}) = (x_j, l_{j+1})$. Note

$$H(l_{j+1}-) - H(x_j) = \frac{1}{n} \sum_{i=1}^{j-1} \frac{H(l_{j+1}-) - H(x_j)}{H(r_i-) - H(l_i)} I(l_{j+1}- \in (l_i, r_i))$$

$$= [H(l_{j+1}-) - H(x_j)] \frac{1}{n} \sum_{i=1}^{j-1} \frac{I(l_{j+1}- \in (l_i, r_i))}{H(r_i-) - H(l_i)}$$

Thus, either $H(l_{j+1}-) - H(x_j) = 0$, or

$$H(l_{j+1}-) - H(x_j) \neq 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{j-1} \frac{I(l_{j+1}- \in (l_i, r_i))}{H(r_i-) - H(l_i)} = 1. \qquad (3.2)$$

Furthermore, it is readily seen that, if (3.2) is true,

$$[H(l_{j+1}-) - H(x_j-)] \geqslant [H(l_{j+1}-) - H(x_j-)] \frac{1}{n} \sum_{i=1}^{j-1} \frac{I(l_{j+1}- \in (l_i, r_i))}{H(r_i-) - H(l_i)} + \frac{1}{n} I(l_{j+1}- \in [r_j, \infty))$$

$$= [H(l_{j+1}-) - H(x_j-)] + \frac{1}{n},$$

which is impossible. Thus $H(l_{j+1}-) - H(x_j) = 0$.

The proofs of (c) and (d) are similar to that of (a) and (b).

Finally consider (e) $(c_m, c_{m+1}) = (r_p, l_j)$, where $l_p < r_p < l_j < r_j$.

Notice that

$$H(r_p) = \frac{1}{n}\sum_{i=1}^{j-1}\frac{H(r_p) - H(l_i)}{H(r_i-) - H(l_i)}I(l_j \in (l_i, r_i)) + \sum_{i=1}^{j-1}\frac{1}{n}I(r_p \in [r_i, \infty))$$

$$H(l_j-) = \frac{1}{n}\sum_{i=1}^{j-1}\frac{H(l_j-) - H(l_i)}{H(r_i-) - H(l_i)}I(l_j - \in (l_i, r_i)) + \sum_{i=1}^{j}\frac{1}{n}I(l_j \in [r_i, \infty))$$

$$= \frac{1}{n}\sum_{i=1}^{j-1}\frac{H(l_j-) - H(l_i)}{H(r_i-) - H(l_i)}I(l_j \in (l_i, r_i)) + \sum_{i=1}^{j}\frac{1}{n}I(r_p \in [r_i, \infty)).$$

Therefore,

$$H(l_j-) - H(r_p) = \frac{1}{n}\sum_{i=1}^{j-1}\frac{H(l_j-) - H(r_p)}{H(r_i-) - H(l_i)}I(l_j \in (l_i, r_i))$$

$$= [H(l_j-) - H(r_p)]\frac{1}{n}\sum_{i=1}^{j-1}\frac{I(l_j \in (l_i, r_i))}{H(r_i-) - H(l_i)}.$$

Thus, either $H(l_j-) - H(r_p) = 0$, or

$$H(l_j-) - H(r_p) \neq 0 \quad and \quad \frac{1}{n}\sum_{i=1}^{j-1}\frac{I(l_j \in (l_i, r_i))}{H(r_i-) - H(l_i)} = 1. \tag{3.3}$$

We now show that case (3.3) leads to a contradiction. Suppose that (3.3) is true. WLOG, we can assume that there is no tie at r_p and at l_j. Then exactly one of the following cases must be true:

(e.1) The point right before r_p, say c_{m-1}, is either l_p or an exact observation, say x_{p_1};

(e.2) c_{m-1} is a left endpoint, say $l_{p_2}, p_2 \neq p$.

If case (e.1) is true, it is easy to reach a contradiction using an argument similar to that of case (a) by considering $H(l_j-) - H(r_p)$ and $H(r_p) - H(c_{m-1})$.

Now suppose that case (e.2) is true. By the definition of H_0, it is easy to see that, for $k \geq 0$, dH_k assigns positive weight to the intervals (l_{p_2}, r_p) and (r_p, l_j). We now prove that the ratio

$$\frac{H_k(l_j) - H_k(r_p)}{H_k(r_p) - H_k(l_{p_2})}$$

is non-increasing in k. In fact,

$$\frac{\dfrac{H_k(l_j) - H_k(r_p)}{H_k(r_i-) - H_k(l_i)}I((r_p, l_j) \subset (l_i, r_i))}{\dfrac{H_k(r_p) - H_k(l_{p_2})}{H_k(r_i-) - H_k(l_i)}I(r_p \in (l_i, r_i])} = \begin{cases} \dfrac{[H_k(l_j) - H_k(r_p)]}{[H_k(r_p) - H_k(l_{p_2})]} & \text{if } (r_p, l_j) \subset (l_i, r_i) \\[2ex] 0 & \text{if } r_p \in (l_i, r_i] \text{ and } l_j \notin (l_i, r_i) \\ & \text{in particular if } i = p \end{cases}$$

$$\frac{H_{k+1}(l_j) - H_{k+1}(r_p)}{H_{k+1}(r_p) - H_{k+1}(l_{p_2})} = \frac{\sum_{i=1}^{j-1} \frac{H_k(l_j) - H_k(r_p)}{H_k(r_i-) - H_k(l_i)} I((r_p, l_j) \subset (l_i, r_i))}{\sum_{i=1}^{j-1} \frac{H_k(r_p) - H_k(l_{p_2})}{H_k(r_i-) - H_k(l_i)} I(r_p \in (l_i, r_i])}$$

$$< \frac{[H_k(l_j) - H_k(r_p)]}{[H_k(r_p) - H_k(l_{p_2})]}$$

and thus $\leq \dfrac{[H_0(l_j) - H_0(r_p)]}{[H_0(r_p) - H_0(l_{p_2})]}.$

Taking limits as $k \to \infty$ yields

$$\frac{H(l_j) - H(r_p)}{H(r_p) - H(l_{p_2})} \leq \frac{[H_0(l_j) - H_0(r_p)]}{[H_0(r_p) - H_0(l_{p_2})]}$$

However, if case (e.2) is true, that is, if $H(l_j) - H(r_p) > 0$ and $H(r_p) - H(l_{p_2}) = 0$, then

$$+\infty = \frac{H(l_j) - H(r_p)}{H(r_p) - H(l_{p_2})} \leq \frac{[H_0(l_j) - H_0(r_p)]}{[H_0(r_p) - H_0(l_{p_2})]} < +\infty$$

The contradiction implies that case (e.2) is impossible. Thus (3.3) is impossible. It follows that $H(l_j) - H(r_p) = 0$. This completes the proof of theorem 2.

Remark 1. The result of theorem 2 does not depend on the choice of the initial distribution H_0, provided that H_0 is chosen by its definition in section 2. Example 2 below indicates that the strictly increasing restriction on H_0 is necessary. The example also shows that a self-consistent estimator is not necessarily a GMLE.

Example 2. Suppose there are only two censoring intervals: (0, 1) and (0.5, 1.5). Let

$$H_0(x) = xI(x \in (0, 0.5)) + \frac{1}{2}I(x \in [0.5, \infty)) + (x - 1)I(x \in [1, 1.5)) + \frac{1}{2}I(x \in [1.5, \infty)).$$

Then $H_k(x) = H_0(x)$ for $k \geq 1$, and thus $H(x) = \lim_{k \to \infty} H_k(x) = H_0(x)$. It is readily seen that non-innermost intervals (0, 0.5) and (1, 1.5) each has weight 1/2, but the innermost interval (0.5, 1) does not have any weight.

The next theorem shows that for each innermost interval the EM and Turnbull's algorithms assign the same weight on it.

Theorem 3
The limiting equation for the EM algorithm

$$H(x) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{H(x) - H(l_i)}{H(r_i-) - H(l_i)} I(x \in (l_i, r_i)) + I(x \in [r_i, \infty)) \right] \tag{3.4}$$

is equivalent to Turnbull's self-consistent equation

$$s_j = \sum_{i=1}^{n} \mu_{ij}(\mathbf{s})/n. \tag{3.5}$$

Proof. Let dH be the measure induced by H and let dS be the measure induced by the self-consistent estimate. It follows from Theorem 2 that both dH and dS assign weight to the

innermost intervals only. Then dH assigns weight $w_j = H(q_j-) - H(p_j)$ $(H(q_j) - H(p_j-))$ to the jth open (closed) innermost interval and dS assigns weight s_j to the jth innermost interval. It suffices to show that w_js satisfy (3.5) and s_js satisfy (3.4).

W.l.o.g., assume that $p_j = l_{j1}$ and $q_j = r_{j2}$, where $j_2 \leqslant j_1$. Note that

(1) $w_j = H(r_{j2}) - H(l_{j1}-)$ if $p_j = q_j$;

(2) $w_j = H(r_{j2}-) - H(l_{j1})$ if $p_j < q_j$.

We first assume $p_j = q_j$, i.e., $r_{j_2} = l_{j_1}$. Then

$$w_j = \frac{1}{n} \sum_{i; (l_i, r_i) \neq [l_{j_1}, r_{j_2}]} \frac{H(l_{j_1}) - H(l_{j_1}-)}{H(r_i-) - H(l_i)} I(l_{j_1} \in (l_i, r_i)) + \frac{1}{n} \sum_{i; [l_i, r_i] = [l_{j_1}, r_{j_2}]} I(l_{j_1} \in [l_i, r_i]),$$

which is the same as

$$w_j = \frac{1}{n} \sum_{i; (l_i, r_i) \neq [l_{j_1}, r_{j_2}]} \frac{w_j}{H(r_i-) - H(l_i)} I(l_{j_1} \in (l_i, r_i)) + \frac{1}{n} \sum_{i; [l_i, r_i] = [l_{j_1}, r_{j_2}]} \frac{w_j}{w_j} I(l_{j_1} \in [l_i, r_i]).$$

It follows that if $p_j = q_j$, then

$$w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{ij} w_j}{\sum_{k=1}^{m} \delta_{ik} w_k}.$$

Similarly, we can show that if $p_j < q_j$, then

$$w_j = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{ij} w_j}{\sum_{k=1}^{m} \delta_{ik} w_k}.$$

This is the same as the equation for s_js, i.e., (3.5).

Analogously we can show that s_js satisfy (3.4).

As mentioned before, the self-consistent GMLE of $F(x)$ is not uniquely defined for $x \in (p_i, q_i)$ if $p_i < q_i$ (Peto, 1973). For the proposed EM algorithm, the value of the estimate $H(x)$ for $x \in (p_i, q_i)$ is uniquely determined once H_0 is determined. It is readily seen that the GMLE defined by (3.4) can be written as a solution of an integral equation as follows.

$$H(x) = \int_0^\infty \int_0^\infty I(l \leqslant x < r) \frac{H(x) - H(l)}{H(r-) - H(l)} dG^*(l, r) + \int_0^\infty \int_0^\infty I(r \leqslant x) dG^*(l, r)$$

$$= \int_0^\infty \int_0^\infty I(l \leqslant x < r) \frac{H(x) - H(l)}{H(r-) - H(l)} dG^*(l, r) + P(R \leqslant x). \tag{3.6}$$

where $G^*(l, r)$ is the distribution function of the observable random vector (L, R). Note that equation (3.6) needs to be modified if we define censoring intervals to be closed $[Y, Z]$ rather than open (Y, Z) as in this paper. Combining theorems 1, 2, and 3, we can prove the following theorem.

Theorem 4
The limiting distribution $H(x)$ of the EM algorithm is the GMLE of F, and is independent of H_0 for $x \notin \mathcal{H}$.

Proof. By theorem 2, the sum of weights on the innermost intervals equals unity, and by theorem 3, the limiting equations of the EM and Turnbull's algorithms are the same. Since Turnbull's algorithm converges to the GMLE, provided that the support of the estimate is on the union of innermost intervals (Turnbull, 1976), the EM algorithm converges to the GMLE, too.

In addition, since the weight of the GMLE on innermost intervals is uniquely determined by the observations given (Peto, 1973), the value of $H(x)$, $x \notin \mathcal{H}$, does not depend on the choice of H_0.

4. Applications

In this section, we shall illustrate the smoothed GMLE technique by using a real data set. It is readily seen that the choice of the initial distribution H_0 does not affect the total amount of weight on innermost intervals but does affect the value of $\hat{S}(x)$ when $x \in (p, q)$, an innermost interval. We present an intuitive approach to choose H_0.

We use the midpoint method. For any $1 \leqslant i \leqslant n$, if $r_i = \infty$, we ignore the interval $[l_i, r_i]$, and if $r_i < \infty$, we let m_i denote the midpoint of $[l_i, r_i]$. Suppose there are k such midpoints and, WLOG, suppose that they are distinct with $m_1 < m_2 < \cdots < m_k$. The initial distribution function H_0 is constructed as follows. Firstly construct an empirical cumulative distribution function (EDF) based on the midpoints $\{m_i, 1 \leqslant i \leqslant k\}$. The EDF jumps at midpoints and is constant between two consecutive midpoints. Secondly, for $1 \leqslant i \leqslant k - 1$, let t_i be the centre point of $[m_i, m_{i+1}]$ and connect points $(t_i, i/k)$ and $(t_{i+1}, (i+1)/k)$ with a line segment. Finally connect $(t_{k-1}, (k-1)/k)$ and $(r^*, 1)$ as well as $(0, 0)$ and $(m_1, 1/k)$ with a line segment, respectively, where $r^* = \max_{1 \leqslant i \leqslant n}\{r_i: r_i < \infty\}$. This constructed polygonal line is the initial distribution H_0 which is continuous and strictly increasing on $[0, r^*]$.

We now use a data set to demonstrate the proposed EM estimator.

Example 3. The following data have been used by Finkelstein & Wolfe (1985) to compare two different treatments for breast cancer patients. The censoring intervals (in months) arose in the follow-up studies for patients treated with radiotherapy and chemotherapy or with radiotherapy alone. The failure time is the time until cosmetic deterioration, as determined by the appearance of breast retraction. The data are reproduced in Tables 1 and 2. The estimate of S for each data set is obtained using the technique derived in this paper. The comparison of the survival functions with the treatments is given in Fig. 1.

Table 1. *Radiotherapy and chemotherapy* (8, 12]

(0, 22]	(24, 31]	(17, 27]	(17, 23]	(24, 30]	(16, 24]	(13, ∞)	
(11, 13]	(16, 20]	(18, 25]	(17, 26]	(32, ∞)	(23, ∞)	(44, 48]	(14, 17]
(0, 5]	(5, 8]	(12, 20]	(11, ∞)	(33, 40]	(31, ∞)	(13, 39]	(19, 32]
(34, ∞)	(13, ∞)	(16, 24]	(35, ∞)	(15, 22]	(11, 17]	(22, 32]	(10, 35]
(30, 34]	(13, ∞)	(10, 17]	(8, 21]	(4, 9]	(11, ∞)	(14, 19]	(4, 8]
(34, ∞)	(30, 36]	(18, 24]	(16, 60]	(35, 39]	(21, ∞)	(11, 20]	(48, ∞)

Table 2. *Radiotherapy alone* (45, ∞)

(6, 10]	(0, 7]	(46, ∞)	(∞)	(7, 16]	(17, ∞)	(7, 14]	
(37, 44]	(0, 8]	(4, 11]	(15, ∞)	(11, 15]	(22, ∞)	(46, ∞)	(46, ∞)
(25, 37]	(46, ∞)	(26, 40]	(46, ∞)	(27, 34]	(36, 44]	(46, ∞)	(36, 48]
(37, ∞)	(40, ∞)	(17, 25]	(46, ∞)	(11, 18]	(38, ∞)	(5, 12]	(37, ∞)
(0, 5]	(18, ∞)	(24, ∞)	(36, ∞)	(5, 11]	(19, 35]	(17, 25]	(24, ∞)
(32, ∞)	(33, ∞)	(19, 26]	(37, ∞)	(34, ∞)	(36, ∞)		
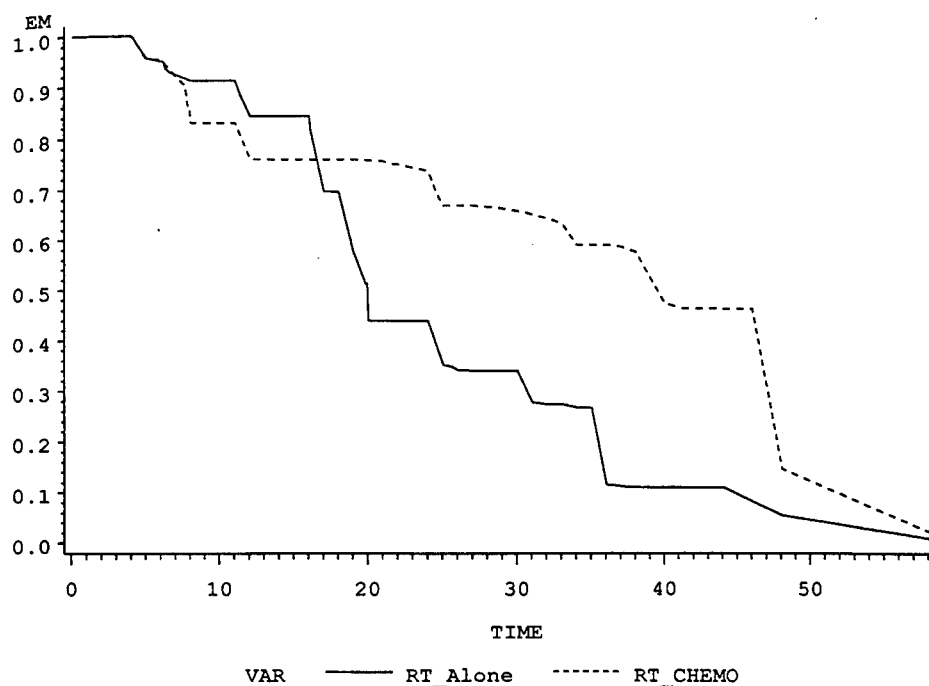
Fig. 1.

References

Becker, N. & Melbye, M. (1991). Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity. *Aust. J. Statist.* **33**, 125–133.

Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.* **18**, 391–404.

Efron, B. (1967). The two sample problem with censored data. *Fourth Berkeley Symposium on Mathematical Statistics.* University of California Press, 831–853.

Finkelstein, D. M. (1986). A proportional hazards model for interval censored failure time data. *Biometrics* **42**, 845–854.

Finkelstein, D. M. & Wolfe, R. A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics* **41**, 933–945.

Groeneboom, P. & Wellner, J. A. (1992). *Information bounds and non-parametric maximum likelihood estimation.* Birkhäuser Verlag, Basel.

Gu, M. G. & Zhang, C-H. (1993). Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.* **21**, 611–624.

Johansen, S. (1978). The product limit estimator as maximum likelihood estimator. *Scand. J. Statist.* **5**, 195–199.

Kaplan, E. L. & Meier, P. (1958). Non-parametric estimation from incomplete observations. *J. Amer. Statist Assoc.* **53**, 457–481.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Statist.*, **27**, 887-906.

Kim, M. Y., De Gruttola, V. G. & Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*, **49**, 13–22.

Peto, R. (1973). Experimental survival curve for interval-censored data. *Appl. Statis.*, **22**, 86–91.

Tsai, W. & Crowley, J. (1985). A large sample study of the generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Statist.*, **13**, 1317–1334.

Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, **38**, 290–295.

Wu, C. F. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.*, **11**, 95–103.

Yu, Q. & Wong, G. (1996a). Estimation of a survival function with interval-censored data. *Submitted for publication.*

Yu, Q. & Wong, G. (1996b). Strong consistency of the generalized MLE of a survival function under the DI model. *Submitted for publication.*

Yu, Q., Li, L. and Wong, G. Y. C. (1997a). On consistency of the self-consistent estimator of survival functions with interval censored data. *Submitted for publication.*

Yu, Q., Li, L. and Wong, G. Y. C. (1997b). Variance of the MLE of a survival function with interval-censored data. *Submitted for publication.*

Linxiong Li, Department of Mathematics, University of New Orleans, New Orleans, LA 70148, USA.

ON CONSISTENCY OF THE SELF-CONSISTENT ESTIMATOR OF SURVIVAL FUNCTIONS WITH INTERVAL-CENSORED DATA

By Qiqing Yu, Linxiong Li and George Y. C. Wong

*Department of Mathematics Sciences, State University of New York
at Binghamton, NY 13902, USA*

*Department of Mathematics, University of New Orleans
New Orleans, LA 70148, USA*

and

*Strang Cancer Prevention Center,
New York, NY 10021, USA*

Abstract The self-consistent estimator is commonly used for estimating a survival function with interval-censored data. Recent studies on interval censoring have focused on case 2 interval censoring, which does not involve exact observations, and double censoring, which involves only exact, right-censored or left-censored observations. In this paper, we consider an interval censoring scheme that involves exact, left-censored, right-censored and strictly interval-censored observations. Under this censoring scheme, we prove that the self-consistent estimator is strongly consistent under certain regularity conditions.

Key words and phrases: Case 2 interval-censored data, exact observations, nonparametric maximum likelihood estimator, self-consistent algorithm, strong consistency.

1. Introduction

Recent studies of interval censoring have focused on case 2 interval-censored (IC) data, which involve a time-to-event variable X whose value is never observed but is known to lie in the time interval between two consecutive inspection times Y and Z. Case 2 interval censoring arises naturally in a longitudinal follow-up study in which the event of interest cannot be easily observed (for instance, cancer recurrence, elevation of levels of a biomaker without any noticeable symptoms).

In this paper, we consider IC data which consist of both case 2 IC data and exact observations. We call such data *mixed IC data*. Mixed IC data do arise in clinical follow-up studies. In a cancer follow-up study in which a tumor marker (for instance, CA 125 in ovarian cancer) is available, a patient whose marker value is consistently on the high (or low) end of the normal range in repeated testing is usually monitored very closely for possible relapse. If such a patient should relapse, then time to clinical relapse can often be accurately determined, and an exact observation is obtained. However, if a patient is not under close surveillance, and would seek help only after some tangible symptoms of the disease have appeared, then time to relapse most likely has to be specified to be within the dates of two successive clinical visits.

Another situation in which such mixed IC data can occur is in the usual right-censored survival analysis where actual dates of events are not recorded, or missing, for a subset of the study population, and can be established only to within specified intervals. An example from the Framingham Heart Study was presented by Odell *et al.* (1992). In this large-scale longitudinal heart disease study, time of occurrence of coronary heart disease (CHD) is recorded for almost every participant. However, time of first occurrence of the CHD subcategory angina pectoris may be specified only as between two clinical visits, several years apart, for some of the participants who suffered from angina pectoris.

For case 2 IC data, Groeneboom and Wellner (1992) proposed the iterative convex minorant algorithm for obtaining the nonparametric MLE (NPMLE) of the distribution function, F, of X. The consistency of the NPMLE and the asymptotic distribution of an alternative estimator are obtained under the assumption that F and the inspection time distribution are both continuous and some additional regularity assumptions. Under the only assumption that the random inspection times are discrete, Yu *et al.* (1998) proved the strong consistency of the NPMLE. They further established the asymptotic normality of the NPMLE by requiring that the inspection times to take on only finitely many values.

Another commonly discussed interval censoring scheme is double censoring. Data are said to be subject to double censoring if they are exact, left censored or right censored; however, they are not to be strictly interval censored. For doubly-censored data, the consistency and asymptotic normality of the self-consistent estimator (SCE) have been established by Turnbull (1974), Chang and Yang (1987), and Gu and Zhang (1993) under different assumptions.

For mixed IC data, Peto (1973) obtained the NPMLE of F using a Newton-Raphson type algorithm. Turnbull (1976) proposed a self-consistent algorithm for estimating F and showed that the associated SCE is also the NPMLE. This SCE has been widely employed in medical applications. See, for example, Finkelstein (1986) and Becker and Melbye (1991). In this paper, we shall establish the strong consistency of the SCE under the assumption that F is arbitrary but the support of the inspection times is finite. Although the NPMLE is consistent with case 2 interval-censored data (Groeneboom and Wellner, 1992), counter example does exist and shows that the SCE may not be consistent with case 2 interval-censored data when the inspection times only take on finitely many values (Yu, 1997). Intuitively, the proof for the consistency of the SCE should be different from that of the NPMLE. We shall show that it is indeed the case in Sections 3 and 4.

The organization of the paper is as follows. Section 2 presents models to describe the mixed IC data and two algorithms for computing the SCE. The strong consistency of the SCE is established in Sections 3 and 4. Some proofs are put in the Appendix.

2. Models For Mixed IC Data

We shall discuss two models for mixed IC data in this section. The one in Section 2.2 is more general than the one in Section 2.1, but we shall show that in terms of the properties of the SCE, it suffices to consider the one in Section 2.1.

2.1. A Simple Model For Mixed IC Data

Let (Y, Z) denote a pair of extended random censoring times (∞ allowed). Assume $Y < Z$ with probability one (w.p.1), and X and (Y, Z) are independent. The observable

mixed IC data are equivalent to a random interval

$$\lfloor L, R \rfloor = \begin{cases} (Y, \infty) & \text{if } Y < X \text{ and } Z = \infty \text{ (right censored)}, \\ (-\infty, Z] & \text{if } X \le Z \text{ and } Y = -\infty \text{ (left censored)}, \\ (Y, Z] & \text{if } -\infty < Y < X \le Z < \infty \text{ (strictly interval censored)}, \\ [X, X] & \text{if } X \notin (Y, Z] \text{ (exact)}. \end{cases} \quad (2.1)$$

Let $\lfloor L_i, R_i \rfloor$, $i = 1, ..., n$, be a random sample from $\lfloor L, R \rfloor$ and $\lfloor l_i, r_i \rfloor$ be a realization of $\lfloor L_i, R_i \rfloor$. Further, let $Q(l, r) = P(L \le l, R \le r)$. Following Peto (1973) and Turnbull (1976), define *sample innermost intervals*, denoted by $\lfloor p_j, q_j \rfloor$'s, to be the nonempty intersections of the intervals $\lfloor l_i, r_i \rfloor$ so that for any pair of intervals $\lfloor p_j, q_j \rfloor$ and $\lfloor l_i, r_i \rfloor$, either $\lfloor p_j, q_j \rfloor \subseteq \lfloor l_i, r_i \rfloor$ or $\lfloor p_j, q_j \rfloor \cap \lfloor l_i, r_i \rfloor = \emptyset$. Note that $\lfloor p_j, q_j \rfloor$ denotes an half open interval if $p_j < q_j$ and a closed interval if $p_j = q_j$. Moreover, every exact observation constitutes an innermost interval. We demonstrate the concept of innermost interval by an example.

EXAMPLE Suppose $n=5$ and the observed intervals $\lfloor l_i, r_i \rfloor$ are $(0, 3], (2, 5], [4, 4], (2, \infty)$, and $(6, 7]$. Then there are three innermost intervals: $(2, 3], [4, 4]$ and $(6, 7]$.

Suppose there are m $(\le n)$ such distinct intervals: $\lfloor p_1, q_1 \rfloor, \lfloor p_2, q_2 \rfloor, ..., \lfloor p_m, q_m \rfloor$, where $p_1 \le q_1 \le p_2 \le q_2 \le \cdots \le q_m$. Define $\delta_{ij} = I(\lfloor p_j, q_j \rfloor \subseteq \lfloor l_i, r_i \rfloor)$, where $I(A)$ denotes the indicator function of the set A.

The self-consistent algorithm (Turnbull, 1976) for obtaining the SCE $\hat{F}_n$ (which assigns weight to innermost intervals only) of F is given by

$$\hat{F}_n(x) = \sum_{j: q_j \le x} s_{nj}, \; x \ge 0,$$

where $\{s_{n1}, ..., s_{nm}\}$ are the probability masses assigned to the corresponding innermost intervals, and satisfy the self-consistent equations

$$s_{nj} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_{ij} s_{nj}}{\sum_{k=1}^{m} \delta_{ik} s_{nk}}, \; j = 1, ..., m. \quad (2.2)$$

Li, Watkins and Yu (1997) proposed an alternative approach based on the EM algorithm for obtaining the SCE $\hat{F}_n$ and expressing $H_n = \hat{F}_n$ as a solution of an integral equation

$$H_n(x) = \int \frac{H_n(x) - H_n(l)}{H_n(r) - H_n(l)} I(l < x < r) dQ_n(l, r) + \frac{1}{n} \sum_{i=1}^{n} I(R_i \le x), \; H_n \in \Theta, \quad (2.3)$$

where $\Theta = \{h: h \text{ is a nondecreasing function from } [-\infty, \infty] \text{ to } [0, 1] \text{ such that } h(-\infty) = 0 \text{ and } h(\infty) = 1\}$ and $Q_n(l, r)$ is the empirical version of $Q(l, r)$. They showed that with proper initial values, algorithms (2.2) and (2.3) give the same weight $s_{nj} = H_n(q_j) - H_n(p_j)$ to the innermost interval $\lfloor p_j, q_j \rfloor$, when it is not closed, or the same weight $H_n(q_j) - H_n(p_j-)$ to $\lfloor p_j, q_j \rfloor$, when it is closed. That is, H_n and $\hat{F}_n$ are equivalent. We shall make use of expression (2.3) to establish the strong consistency of $\hat{F}_n$ in Sections 3 and 4.

3

Following the identifiability assumption given in Chang and Yang (1987), we define $K(x) = P\{X$ is not censored $|X = x\}$ for each x. Let $\tau_l = \inf\{x : K(x) = 0\}$ and $\tau_r = \sup\{x : K(x) = 0\}$, if $\{x : K(x) = 0\} \neq \emptyset$. Otherwise, define $\tau_l = \tau_r = \infty$. For each $x \in (\tau_l, \tau_r)$, either $K(x) = 0$ or $K(x)$ is not defined. To see this, it suffices to show that for any two points $a < b$ satisfying $K(a) = K(b) = 0$, there do not exist $x \in (a, b)$ such that $K(x) > 0$. In fact, $K(a) = 0$ implies that $P\{Y < a\} \geq P\{Y < a \leq Z\} = 1$. Also, $K(b) = 0$ implies that $P\{Z \geq b\} \geq P\{Y < b \leq Z\} = 1$. Thus

$$P\{Y \leq \tau_l \leq \tau_r \leq Z\} = 1 \text{ and } K(x) = 0 \ \forall \ x \in (\tau_l, \tau_r). \tag{2.4}$$

There are only four possible cases that model (2.1) implies: (1) $\tau_l > -\infty$ and $\tau_r = \infty$, (2) $\tau_l = -\infty$ and $\tau_r < \infty$, (3) $-\infty < \tau_l < \tau_r < \infty$, and (4) $\{x : K(x) = 0\} = \emptyset$. Case (1) is a right censorship model as $P(Z = \infty) = 1$ by (2.4). Moreover, case (2) is a left censorship model. Both of them do not allow strictly interval-censored observations. If case (3) is true, then $Y \leq \tau_l$ and $Z \geq \tau_r$ w.p.1, which is not practically realistic. Thus case (4) is the only practical case in model (2.1) that includes both strictly interval-censored observations and exact observations. In next subsection we see how to extend model (2.1) to cover more general situations.

2.2. A Model for More General Mixed IC Data.

Even though case (4) in Section 2.1 does not have the drawback as in the first three cases, it implies that $P\{K(X) > 0\} = 1$. It is often the case that a study can only last for a certain period of time, say, a time interval $[a, b]$, where $0 \leq F(a) < F(b) < 1$. In such a case, the mixed interval-censored observation $\lfloor L, R \rfloor$ satisfies

$$\{L \text{ or } R \in (-\infty, a) \cup (b, +\infty)\} = \emptyset. \tag{2.5}$$

Consequently, $P\{K(X) > 0\} \leq P\{a < X \leq b\} < 1$. Thus, model (2.1) cannot specify such mixed IC data. Note that (2.1) is equivalent to

$$\lfloor L, R \rfloor = \begin{cases} (Y, Z] & \text{if } X \in (Y, Z], \\ [X, X] & \text{if } X \notin (Y, Z]. \end{cases} \tag{2.6}$$

We now formulate a model for mixed IC data satisfying (2.5). Assume $Y < Z$ w.p.1., and $\{Y \text{ or } Z \in (-\infty, a) \cup (b, \infty)\} = \emptyset$. Suppose that X and (Y, Z) are independent and the observable random vector

$$\lfloor L, R \rfloor = \begin{cases} (Y, Z] & \text{if } X \in (Y, Z], \\ [X, X] & \text{if } X \notin (Y, Z] \text{ and } a < X \leq b, \\ (-\infty, a] & \text{if } X \notin (Y, Z] \text{ and } X \leq a, \\ (b, \infty) & \text{if } X \notin (Y, Z] \text{ and } X > b. \end{cases} \tag{2.7}$$

In the case of (2.5) or (2.7), we can only estimate $F(x)$ for x in $[a, b]$, or equivalently, the cdf F^* of X^*, where $X^* = aI(X \leq a) + XI(a < X \leq b) + 2bI(X > b)$. Note that X^* and (Y, Z) are independent. Due to (2.5) or (2.7), the right-censored observation (b, ∞) will always be an innermost interval. The NPMLE (or an SCE) $\hat{F}(x)$ is not uniquely

determined for $x \in (b, \infty)$ (see, e.g., Peto, 1973), though the total mass assigned by the NPMLE (or the SCE) to the interval (b, ∞) is uniquely determined. Thus we can, without loss of generality (WLOG), assume that the mass is put at the point $2b$ ($\in (b, \infty)$). In other words, (b, ∞) can be treated as an exact observation $[2b, 2b]$. For a similar reason, the left-censored observation $(-\infty, a)$ can be treated as an exact observation $[a, a]$. Thus model (2.7) is equivalent to

$$
\lfloor L, R \rfloor = \begin{cases} (Y, Z] & \text{if } X^* \in (Y, Z], \\ [X^*, X^*] & \text{if } X^* \notin (Y, Z], \end{cases} \tag{2.8}
$$

If $F(a) = 0$ and $F(b) = 1$, then models (2.7) and (2.8) are the same as (2.1) (or (2.6)).

In view of (2.6) and (2.8), it is easy to see that in the case of (2.5) or (2.7), in order to estimate F, it suffices to estimate F^*, which reduces model (2.7) to model (2.1). Similar modification can be made to handle the situation that there are no observations L or R in a union of arbitrary intervals. In view of the above discussion, we shall focus on model (2.1) for the rest of the paper.

3. Consistency In Case Of Finite Support For F

In this section, we assume that both the support of X, say S_F, and the support of Y and Z, say S_G, contain finitely many points. The generalization of F to an arbitrary distribution function is given in Section 4. The assumption concerning S_G is a reasonable one. In practice inspections of most follow-up studies are recorded on a discrete time scale (daily, weekly, monthly, etc.), and the total study period is finite, so the number of censoring points, i.e. the support of Y and Z, is also finite. Such an assumption was adopted by Finkelstein (1986) and Becker and Melbye (1991), among others.

Suppose that X takes on values $x_1, x_2, \ldots, x_u$, and $\lfloor L, R \rfloor$ takes on values $I_1 = \lfloor l_1^o, r_1^o \rfloor$, $I_2 = \lfloor l_2^o, r_2^o \rfloor$, $\ldots, I_v = \lfloor l_v^o, r_v^o \rfloor$ with probability $e_i = P\{L = l_i^o, R = r_i^o\} > 0$. Based on the assumption that $K(x) > 0$ for all $x > 0$, Chang and Yang (1987) and Gu and Zhang (1993) proved the consistency of the SCE for doubly-censored observations. In this paper, we weaken this assumption and prove the consistency of the SCE on the set $\mathcal{O} = \{x; K(x) > 0\}$ with mixed IC data.

For a point x satisfying $K(x) = 0$ and $P\{X = x\} > 0$, since there are no exact observations available at this point, the distribution function F is not estimable, and hence consistency cannot be assessed. Let us consider the structure of the innermost intervals as sample size $n \to \infty$. For x_i, if $K(x_i) > 0$, then it follows from the strong law of large number that $P\{X_k \neq x_i, \text{ for all } k = 1, 2, \ldots, n\} \to 0$ as $n \to \infty$. In other words, $K(x_i) > 0$ implies that $[x_i, x_i]$ is an innermost interval w.p.1, which further implies that the union of all closed innermost intervals coincides with the set $\mathcal{O}$. Let $A_1, A_2, \ldots, A_{m_1}$ be the innermost intervals induced by the intervals $I_i, i = 1, 2, \ldots, v$, and call them *population innermost intervals*. It is seen that as $n \to \infty$ the set of sample innermost intervals induced by $\lfloor L_i, R_i \rfloor, i = 1, 2, \ldots, n$ converges almost surely to the set of population innermost intervals. Since we are only concerned with large sample properties, we can, WLOG, assume that the sample size is large enough so that $m = m_1$. For the rest of the paper, m will be used to denote both the number of population innermost intervals and the number of sample innermost intervals. Also we shall suppress the qualifier w.p.1 throughout the rest of the paper to avoid repetition.

Let $\mathbf{s}_n = (s_{n1}, s_{n2}, \ldots, s_{nm})$ be a solution of (2.2). For sufficiently large n, the s_{ni}'s are the masses assigned to the innermost intervals by the SCE. Since $\{H_n, n \geq 1\}$ is a bounded monotone sequence, it follows from Helly-Bray selection theorem that there exist a subsequence, say $\{n_k\}$, of integers, a function H and a vector $\mathbf{s}$, such that $\lim_{n_k \to \infty} H_{n_k}(x) = H(x)$ and $\lim_{n_k \to \infty} \mathbf{s}_{n_k} = \mathbf{s} = (s_1, s_2, \ldots, s_m)$. Taking the limit in (2.2) and (2.3) with respect to n_k, we obtain

$$s_j = \sum_{i=1}^{v} e_i \frac{\delta_{ij} s_j}{\sum_{k=1}^{m} \delta_{ik} s_k}, \quad s_j \geq 0, \ j = 1, \ldots, m, \quad \sum_{j=1}^{m} s_j = 1, \tag{3.1}$$

where $\delta_{ij} = I(A_j \subseteq \lfloor l_i^o, r_i^o \rfloor)$, and

$$H(x) = \int_{l \leq x < r} \frac{H(x) - H(l)}{H(r) - H(l)} dQ(l, r) + P\{R \leq x\}, \quad H \in \Theta, \tag{3.2}$$

since Q_{n_k} converges to Q almost surely as $n_k \to \infty$.

We state two lemmas. The proof of Lemma 1 is relegated to the appendix. Hereafter, the discussion regarding uniqueness of the solution $H(x)$ of Eq. (3.2) will be restricted to the set $\mathcal{O}$.

Lemma 1 *Let $\mathbf{s}^o = (s_1^o, s_2^o, \ldots, s_m^o)$ where $s_j^o = P\{X \in A_j\}$, $j = 1, 2, \ldots, m$. Then $\mathbf{s} = \mathbf{s}^o$ is the unique solution of Eq. (3.1).*

Lemma 2 (Li, Watkins and Yu, 1997) *Let dH be the measure induced by a c.d.f. H and $s_j = dH(A_j)$ for all j. Then $\mathbf{s} = (s_1, \ldots, s_m)$ is a solution to Eq. (3.1) if and only if H is a solution to Eq. (3.2).*

Theorem 1 *Suppose S_F and S_G contain finitely many points. Then (1) F is the unique solution of Eq. (3.2) for $x \in \mathcal{O}$, and (2) the SCE $H_n(x)$ of $F(x)$ satisfies $\sup_{x \in \mathcal{O}} |H_n(x) - F(x)| \to 0$, as $n \to \infty$.*

Proof. Since $s_j^o = dF(A_j)$, F is a solution of Eq. (3.2) by Lemmas 1 and 2. Meanwhile, for each solution H of (3.2), dH is uniquely determined by Lemmas 1 and 2 again. Consequently, statement (1) follows.

It follows from the convergence of H_{n_k} and (1) above that $H_{n_k}(x) \to F(x)$ as $n_k \to \infty$ for $x \in \mathcal{O}$. The convergence of $H_n(x)$ to $F(x)$ for $x \in \mathcal{O}$ follows from Helly-Bray selection theorem. The uniform convergence of H_n is immediate by the assumption that S_F and S_G contain only finitely many points. $\square$

REMARK. Even though $F(x)$ is not estimable for $x \in (\tau_l, \tau_r)$, it is easy to see that $H(\tau_r) - H(\tau_l) = F(\tau_r) - F(\tau_l)$. This remark is also valid for model (2.7). Moreover, under model (2.7) $F(x)$ is not estimatable for $x < a$ or $x > b$.

4. Consistency Of H_n For Arbitrary F

6

In this section, we extend the result of the previous section to the case where G is the same as previously defined but F is arbitrary.

Theorem 2 *Suppose that (Y, Z) takes on finitely many values and F is arbitrary. Then the solution to Eq. (3.2) is unique. Furthermore, $H_n(x)$ converges to $F(x)$ uniformly for all $x \in \mathcal{O}$.*

Proof. The main idea of the proof is to partition the interval $[0, \infty)$ into finitely many subintervals, and then to prove the consistency of the SCE for every such subinterval.

WLOG, assume that the values that (Y, Z) can take are (a_i, b_i), $i = 1, ..., N_G$, for some integer N_G. Rank the $2N_G$ values $\{a_i, b_i\}$ in increasing order to obtain a sequence (ties and infinity are allowed). Let $d_1 < d_2 < \cdots < d_N$ ($N \leq 2N_G$) be the distinct finite values of the sequence. We first partition $[0, \infty)$ into

$$[0, 0], (0, d_1), [d_1, d_1], (d_1, d_2), \ldots, [d_N, d_N], (d_N, \infty). \qquad (4.1)$$

Note that in this partition, all exact observations in the same interval (d_j, d_{j+1}) (or $[d_j, d_j]$) carry the same weight. This is because for any observed interval, if (d_j, d_{j+1}) (or $[d_j, d_j]$) is not a subset of the interval, then it is disjoint from the observed interval, and because the weight received by an innermost interval is determined by all the observed intervals that cover the innermost interval (see (2.2)).

For a fixed $\epsilon > 0$, if there is a value d in an open interval (d_j, d_{j+1}) such that $P\{X = d\} \geq \epsilon$, divide the interval into $(d_j, d), [d, d], (d, d_{j+1})$. Perform the partitioning for every such d. Since the set of such d values is finite, the total number of intervals partitioning $[0, \infty)$ must also be finite. WLOG, assume (4.1) is the final partition at this stage.

Consider an interval, say, (d_1, d_2), such that $F(d_2-) - F(d_1) > 0$. For this fixed ϵ, partition (d_1, d_2) into subintervals, say $(c_1, c_2), [c_2, c_2], (c_2, c_3,), \ldots, (c_k, c_{k+1})$ ($c_1 = d_1$ and $c_{k+1} = d_2$) such that $F(c_{i+1}-) - F(c_i) < \epsilon$ for $i = 1, 2, \ldots, k$. Perform this second partition for every interval (d_i, d_{i+1}) and $[d_i, d_i]$ for all $i = 0, 1, \ldots, N$, where $d_0 = 0$.

From now on, we focus our discussion on (d_1, d_2). The argument for other intervals is similar. Let c_i' be the midpoint of the interval (c_i, c_{i+1}), $i = 1, ..., k$, and construct a new (pseudo) distribution function F' with finite support, $F'(d_i) = F(d_i)$ and $F'(c_i') - F'(c_i'-) = F(c_{i+1}-) - F(c_i)$, for all i. It is readily seen that

$$\sup_x |F(x) - F'(x)| < \epsilon. \qquad (4.2)$$

It can be verified that if (τ_l, τ_r) is not an empty set, then one of (d_i, d_{i+1}) must be (τ_l, τ_r), due to the special structure of partition (4.1). In addition, since consistency is restricted to $\mathcal{O}$, it is natural to assume that $S_F \cap (d_1, d_2) \subset \mathcal{O}$. Moreover, since $F(d_2-) - F(d_1) > 0$ and $S_F \cap (d_1, d_2) \subset \mathcal{O}$, the probability of having exact observations in (d_1, d_2) converges to one as $n \to \infty$. Thus, for n large enough, we can eventually observe exact observations in the interval and hence $(d_1, d_2]$ cannot be an half open innermost interval.

Consider a pseudo random variable $X' = \sum_i [c_i' I(X \in (c_i, c_{i+1}) + c_i I(X = c_i)]$. Note that X' has the distribution function F'. Suppose there are h exact observations in (c_i, c_{i+1}), then the pseudo random variable X' will assume the value c_i' as an exact

observation a total of h times. For sample size n, let w_{ni} denote the weight received by all exact observations in (c_i, c_{i+1}), and let $w'_{ni} = w_{ni}$ be the weight received by c'_i from the pseudo observations generated by X'. Let H'_n denote the corresponding SCE of F' associated with X' and H_n the SCE of F associated with X. It is easy to see that for each i and $D_i = (d_i, d_{i+1})$ or $[d_i, d_i]$,

$$dH_n(D_i) = dH'_n(D_i). \tag{4.3}$$

By the results of the previous section and the finiteness of support of F', $H'_n(d_i) \to F'(d_i)$ for each i as $n \to \infty$. Thus, it follows from (4.2) and (4.3) that when n is large enough,

$$\sup_{x \in \mathcal{O}} |F(x) - H_n(x)| \le \epsilon, \tag{4.4}$$

which proves the consistency of the SCE H_n. By Helly-Bray selection theorem and the fact that F is a solution of (3.2), (4.4) also shows that F is the unique solution of (3.2). $\square$

Appendix

In this appendix, we give a proof of Lemma 1. To show the uniqueness of the solution of (3.1), consider the generalized log-likelihood function

$$\Lambda = \Lambda(\mathbf{s}) = \sum_{i=1}^{v} e_i \left(\ln \sum_{j=1}^{m} \delta_{ij} s_j \right).$$

It follows from (2.2) that $\sum_{j=1}^{m} \delta_{ij} s_j > 0$ for every i, $1 \le i \le v$ and hence the function Λ is well defined. Let $\mathbf{s}^e = (s_1^e, s_2^e, \ldots, s_m^e)$ denote the solution of Eq. (3.1) and $\mathbf{s}^o = (s_1^o, s_2^o, \ldots, s_m^o)$ the probability vector with $s_i^o = \mathrm{P}\{X \in A_i\}$.

To facilitate the proof of Lemma 1, we first establish three lemmas. Following the notations of Section 3, let $I_i = \lfloor l_i^o, r_i^o \rfloor$ and $e_i = P\{\lfloor L, R \rfloor = I_i\} = \alpha_i^o g_i$, where

$$\alpha_i^o = P\{X \in I_i\} = \sum_{j=1}^{m} \delta_{ij} s_j^o, \quad i = 1, \ldots, v, \tag{A.1}$$

and $g_i = P\{\lfloor Y, Z \rfloor = I_i\}$ if $l_i^o < r_i^o$ and $P\{l_i^o \notin (Y, Z)\}$ if $l_i^o = r_i^o$. Thus $\sum_i \alpha_i^o g_i = \sum_i e_i = 1$. It is seen that $\mathbf{s}^o$ uniquely determines $\mathbf{a}^o = (\alpha_1^o, \alpha_2^o, \ldots, \alpha_v^o)$. Let $\mathbf{a} = (\alpha_1, \ldots, \alpha_v)$, with $\alpha_i = \sum_{j=1}^{m} \delta_{ij} s_j$. Then Λ can be rewritten as

$$\Lambda(\mathbf{s}) = \sum_{i=1}^{v} e_i \ln \alpha_i \equiv \lambda(\mathbf{a}).$$

Thus maximizing $\Lambda(\mathbf{s})$ is equivalent to maximizing $\lambda(\mathbf{a})$. Note that α_i^o, e_i and g_i are fixed, but α_i are not.

8

Let X_s be a random variable such that $P\{X_s \in A_j\} = s_j$, $j \geq 1$. Define a new random interval $\lfloor L_s, R_s \rfloor$ to be the counterpart of $\lfloor L, R \rfloor$ in (2.1) with X replaced by X_s. Thus α_i satisfies $\alpha_i g_i = P\{\lfloor L_s, R_s \rfloor = I_i\}$. It follows that

$$\sum_i \alpha_i g_i = 1. \tag{A.2}$$

Lemma A1 *Suppose* $\mathbf{a}$ *satisfies (A.2). Then* $\lambda(\mathbf{a})$ *is uniquely maximized by* $\mathbf{a}^o$, *where* $\alpha_i^o = e_i/g_i$.

Proof. Let $t_i = \alpha_i g_i$, $i = 1, ..., v$. Then $t_i \in [0, 1]$ and $\sum_i t_i = 1$. Let $h(\mathbf{t}) = \sum_{i=1}^v e_i \ln t_i$. Note that

$$h(\mathbf{t}) = \sum_i e_i \ln \alpha_i + \sum_i e_i \ln g_i = \lambda(\alpha) + \sum_i e_i \ln g_i$$

and $\sum_i e_i \ln g_i$ is fixed under the given assumption. Hence, maximizing $\lambda(\mathbf{a})$ is equivalent to maximizing $h(\mathbf{t})$. It can be shown that $h(\mathbf{t})$ is uniquely maximized by $t_i = e_i$, $i \geq 1$. Therefore, the unique maximizer for $\lambda(\mathbf{a})$ is $\alpha_i^o = e_i/g_i$. $\square$

Lemma A2 $\mathbf{s}^o$ *is the unique maximum point of* $\Lambda(\mathbf{s})$.

Proof. Following the notations in Lemma A1, we have $\Lambda(\mathbf{s}) = \lambda((\mathbf{a}(\mathbf{s})))$. By Lemma A1 and the equality $\mathbf{a}(\mathbf{s}^o) = \mathbf{a}^o$ (due to (A.1)), $\mathbf{s}^o$ is a maximum point of $\Lambda(\mathbf{s})$. By the finiteness assumption on S_F and S_G, each population innermost interval is a realization of $\lfloor L, R \rfloor$, say, $A_j = I_{i_j}$, except perhaps (τ_l, τ_r) if (τ_l, τ_r) is not an empty set. Thus (A.1) implies that there are at least $m - 1$ (out of v) js such that $\alpha_{i_j}^o = s_j^o$. Since $\lambda((\mathbf{a}(\mathbf{s})))$ is uniquely maximized by $\mathbf{a}^o$, $\mathbf{s}^o$ is the unique maximum point of $\Lambda(\mathbf{s})$. $\square$

Lemma A3 $s_k^o > 0$ *implies that* $s_k^e > 0$.

Proof. Note that for each k, if $P\{X \in A_k\} > 0$, then there is an integer h such that $I_h = A_k$ and thus $\delta_{hk} = 0$ if $h \neq k$ and 1 otherwise. For $j = k$, Eq. (3.1) yields

$$s_k^e \geq e_h \frac{s_k^e}{s_k^e} = e_h > 0 \qquad (\text{since } I_h = A_k \text{ and } e_h > 0).$$

This completes the proof of the lemma. $\square$

We are now ready to prove Lemma 1. By Lemma A2, $\mathbf{s}^o$ is the unique maximizer of Λ. Thus, to prove the lemma, it suffices to show that $\mathbf{s}^o = \mathbf{s}^e$. Consider the effect of increasing a particular component, s_k, by a small amount u and then dividing all the s_j, including $s_k + u$, by $1 + u$ in order to ensure that the components of $\mathbf{s}$ sum to 1. Let $d_k(\mathbf{s}) = \frac{\partial \Lambda}{\partial u}\big|_{u=0}$. Then

$$d_k(\mathbf{s}) = \frac{\partial}{\partial u} \Lambda\left(\frac{s_1}{1+u}, ..., \frac{s_k + u}{1+u}, ..., \frac{s_m}{1+u}\right)\Big|_{u=0}$$

9

$$= \sum_{i=1}^{v} e_i \frac{\partial}{\partial u} \ln \sum_{j=1}^{m} \delta_{ij} s_j^r \Big|_{u=0} \qquad \left(s_j^r = \begin{cases} s_j/(1+u) & \text{if } j \neq k \\ (s_k+u)/(1+u) & \text{if } j = k \end{cases} \right)$$

$$= \sum_{i=1}^{v} e_i \frac{\sum_{j=1}^{m} \delta_{ij} \frac{\partial}{\partial u} s_j^r}{\sum_{j=1}^{m} \delta_{ij} s_j^r} \Big|_{u=0}$$

$$= -\sum_{i=1}^{v} e_i \left(1 - \frac{\delta_{ik}}{\sum_{j=1}^{m} \delta_{ij} s_j} \right), \quad k = 1, ..., m. \tag{A.3}$$

Consider two separate situations regarding the values of s_k^o.

CASE 1. $s_k^o > 0$, for all k.

If $\mathbf{s}^e$ is a solution to (3.1), then it follows that $s_k^e > 0$ for all k. Consequently,

$$0 = \sum_{i=1}^{v} e_i \left(1 - \frac{\delta_{ik}}{\sum_{j=1}^{m} \delta_{ij} s_j^e} \right), \quad k = 1, ..., m, \tag{A.4}$$

since $\sum_{i=1}^{v} e_i = 1$. In view of (A.3) and (A.4), $d_k(\mathbf{s}^e) = 0$ and $s_k^e > 0$ for each k. Therefore, $\mathbf{s}^e$ is the maximum point of L. By Lemma A2, $\mathbf{s}^e = \mathbf{s}^o$.

CASE 2. $s_k^o = 0$ for some k.

WLOG, assume that $s_m^o = 0$ and $s_k^o > 0$ for $k = 1, 2, \ldots, m-1$. We shall show that $s_m^e > 0$ leads to a contradiction. If $s_m^e > 0$, then $s_i^e > 0$ for all i by Lemma A3. Consequently, (A.4) holds. By virtue of (A.3) and (A.4), $\mathbf{s}^e$ is a maximum point of Λ, and it follows that $\mathbf{s}^e = \mathbf{s}^o$. However, this contradicts the hypothesis that $s_m^e > s_m^o = 0$. This completes the proof for Case 2 and thus the proof of the lemma. □

References

[1] Becker, N. and Melbye, M. (1991). Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity. *Austral. J. Statist.*, 33, 125-133.

[2] Chang, M.N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly-censored data. *Ann. Statist.* 15. 1536-1547.

[3] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42, 845-854.

[4] Groeneboom, P. and Wellner, J. A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.*

[5] Gu, M.G. and Zhang, C-H. (1993). Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.*, 21, 611-624.

[6] Li, L., Watkins, T. and Yu, Q. (1997). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval - censored data. *Scand. J. Statist.*, 24, 531-542.

[7] Odell, P.M., Anderson, K.M. and D'Agostino, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics*, 48, 951-959.

[8] Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, 22, 86-91.

[9] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *JASA*, 69, 169-173.

[10] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B.* 38, 290-295.

[11] Yu, Q., Schick, A., Li, L. and Wong G. (1998). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Statistics & Probability Letters*, 23, 223-228.

[12] Yu, Q. (1997). Nonparametric and parametric estimation with interval-censored data. *Technical report*. Department of Mathematical Sciences, State University of New York at Binghamton.

Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times*

Qiqing YU, Anton SCHICK, Linxiong LI and George Y.C. WONG

State University of New York at Binghamton (Q.Y., A.S.), *University of New Orleans* (L.L.) and *Strang Cancer Prevention Center* (G.Y.C.W.)

ABSTRACT

We consider the case 1 interval censorship model in which the survival time has an arbitrary distribution function F_0 and the inspection time has a discrete distribution function G. In such a model one is only able to observe the inspection time and whether the value of the survival time lies before or after the inspection time. We prove the strong consistency of the generalized maximum-likelihood estimate (GMLE) of the distribution function F_0 at the support points of G and its asymptotic normality and efficiency at what we call regular points. We also present a consistent estimate of the asymptotic variance at these points. The first result implies uniform strong consistency on $[0, \infty)$ if F_0 is continuous and the support of G is dense in $[0, \infty)$. For arbitrary F_0 and G, Peto (1973) and Turnbull (1976) conjectured that the convergence for the GMLE is at the usual parametric rate $n^{\frac{1}{2}}$. Our asymptotic normality result supports their conjecture under our assumptions. But their conjecture was disproved by Groeneboom and Wellner (1992), who obtained the nonparametric rate $n^{\frac{1}{3}}$ under smoothness assumptions on the F_0 and G.

RÉSUMÉ

Nous considérons le modèle de censure d'intervalle de cas 1 dans lequel le temps de survie a une fonction de répartition arbitraire F_0 et le temps d'inspection a une fonction de répartition discrète G. Dans un tel modèle on est seulement capable d'observer le temps d'inspection et si la valeur du temps de survie est superieure ou inférieure le temps d'inspection. Nous prouvons convergence forte de l'estimateur du maximum de vraisemblance généralisé (GMLE) de la fonction de répartition F_0 aux points de support de G et sa normalité asymptotique et l'efficacité à ce que l'on appelle les points réguliers. Nous présentons également un estimateur convergent de la variance asymptotique à ces points. Le premier résultat implique une convergence uniforme forte sur $[0, \infty)$ si F_0 est continu et le support de G est dense en $[0, \infty)$. Pour des F_0 et G arbitraires, Peto (1973) et Turnbull (1976) ont conjecturé que la convergence du GMLE est au taux paramétrique habituel de $n^{\frac{1}{2}}$. Notre résultat de normalité asymptotique supporte leur conjecture sous nos hypothèses. Mais leur conjecture a été réfutée par Groeneboom et Wellner (1992) qui ont obtenu le taux non-paramétrique de $n^{\frac{1}{3}}$ sous des hypothèse de F_0 et G lisses.

1. INTRODUCTION

In survival analysis, one frequently is unable to precisely observe the survival time X

of interest, but can only assess that it belongs to some random interval. The simplest such model is the so-called case 1 interval-censorship model. In this model one is only able to observe a random time Y and whether x lies in the random interval $[0, Y]$ or (Y, ∞). More formally, one observes (Y, Δ), where $\Delta = I[X \leq Y]$. Here and below $I[A]$ denotes the indicator function of the event A. The random time Y is called the *inspection time*.

Such data arise in industrial life testing and medical research. Consider for example an animal sacrifice study in which a laboratory animal has to be dissected to check whether a tumour has developed. In this case, X is the onset of tumour and Y is the time of the dissection, and we only can infer at the time of dissection whether the tumour is present or has not yet developed. Other examples are mentioned in Ayer *et al.* (1955), Keiding (1991) and Wang and Gardiner (1996).

We shall assume throughout that the lifetime X and the inspection time Y are independent and denote their distribution functions by F_0 and G, respectively. Our data consist of n independent copies $(Y_i, \Delta_i) = (Y_i, I[X_i \leq Y_i])$, $i = 1, \ldots, n$, of (Y, Δ). We consider estimating (characteristics of) the distribution function F_0 based on these data.

Ayer *et al.* (1955) derived the explicit expression of the generalized maximum-likelihood estimator (GMLE) of the distribution function F_0. Moreover, they established the weak consistency of the GMLE at continuity points x of F_0 under additional assumptions on G. They also mentioned the strong consistency of the GMLE at each support point of a discrete Y with finitely many values. Using an inequality of theirs, we shall generalize this result to arbitrary discrete Y in our Theorem 2.1. From this result we shall derive the uniform strong consistency on the entire line if F_0 is continuous and the support of Y is dense in the positive half line. Moreover, using Theorem 2.1 of Ayer *et al.* (1955), we shall derive another explicit representation of the GMLE at what we call regular points and conclude with its aid the asymptotic normality and efficiency of the GMLE at such points.

Peto (1973) considered the problem of obtaining the GMLE based on interval-censored data using a Newton–Raphson algorithm. Turnbull (1976) proposed a self-consistent algorithm and showed that it converges to the GMLE $\hat{F}$. Both conjectured that for arbitrary F_0 and G, the GMLE is asymptotically normal at the usual $n^{\frac{1}{2}}$ rate. Thus our results provide a partial justification of their claim for discrete Y. It was, however, shown by Groeneboom and Wellner (1992) that this conjecture is false if F_0 and G satisfy certain smoothness assumptions. Indeed, their Theorem 5.1 establishes that under differentiability assumptions on F_0 and G the convergence is at the slower $n^{\frac{1}{3}}$ rate and the limiting distribution is not normal. Groeneboom and Wellner (1992) also obtained the uniform strong consistency of the GMLE for continuous F_0 and G. A variant of this result was also proved by Wang and Gardiner (1996) using a totally different approach and a slightly different set of assumptions.

The results of Groeneboom and Wellner (1992) give a fairly detailed description for the case of continuous F_0 and G, while ours do so for the case of arbitrary F_0 and discrete G. There are many practical situations in which Y is discrete. In medical research, for example, the data are often recorded as integers (to represent number of days, weeks etc.). Motivated by this, we assume that the inspection time Y is a discrete random variable with density g. This assumption is used by several authors in survival analysis: Becker and Melbye (1991) and Finkelstein (1986) among others.

Our paper is organized as follows. We introduce the GMLE in Section 2 and prove its strong consistency. In Section 3 we establish the asymptotic normality and efficiency of the GMLE at what we call regular points. Finally, Section 4 summarizes our work, discusses some of its implications, addresses some questions raised by it and establishes

connections with the work of others. In particular, we show by means of an example that our asymptotic normality result fails at nonregular points even though the rate of convergence is still $n^{\frac{1}{2}}$.

2. THE CONSISTENCY OF THE GMLE

By our assumptions, Y is a discrete random variable with density g. Let $\mathcal{A}$ be the set of possible values of Y, i.e., $\mathcal{A} = \{a \in \mathbb{R} : g(a) > 0\}$. For $a \in \mathcal{A}$, set

$$N_n^-(a) = \frac{1}{n} \sum_{j=1}^{n} I[X_j \leq a, \ Y_j = a],$$

$$N_n^+(a) = \frac{1}{n} \sum_{j=1}^{n} I[X_j > a, \ Y_j = a],$$

$$N_n(a) = \frac{1}{n} \sum_{j=1}^{n} I[Y_j = a].$$

The generalized likelihood is given by

$$\Lambda_n(F) = \prod_{a \in \mathcal{A}} F(a)^{nN_n^-(a)} \{1 - F(a)\}^{nN_n^+(a)}.$$

In the above we let F range over the set $\mathcal{F}$ of all subdistribution functions. A function F is called a *subdistribution function* if $F = aF_1$ for some distribution function F_1 and some number a in $[0, 1]$. Thus a subdistribution function has all the properties of a distribution function except that its limit at infinity may be less than 1.

Note that $\Lambda_n(F)$ depends on F only through the values of F at the points $a \in \mathcal{A}$ for which $N_n(a) > 0$. Thus there exists no unique maximizer of $\Lambda_n(F)$ in the set $\mathcal{F}$. But there exists a uniquely determined $\mathcal{F}$-valued random element $\hat{F}_n$ which maximizes $\Lambda_n(F)$ and satisfies $\hat{F}_n(b) = \sup\{\hat{F}_n(a) : a \leq b, N_n(a) > 0\}$ for each $b \in \mathbb{R}$. Here we interpret the supremum of the empty set as 0. We call $\hat{F}_n$ the GMLE of F_0. It is easy to check that $\hat{F}_n(Y_{(1)}) = 0$ on the event $\{N_n^-(Y_{(1)}) = 0\}$ and $\hat{F}_n(Y_{(n)}) = 1$ on the event $\{N_n^+(Y_{(n)}) = 0\}$, where $Y_{(1)}$ and $Y_{(n)}$ are the smallest and largest among $Y_1, \ldots, Y_n$. For latter use, set

$$\tilde{F}_n(a) = \begin{cases} N_n^-(a)/N_n(a) & \text{if} \quad N_n(a) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

THEOREM 2.1. *The GMLE $\hat{F}_n$ satisfies $\hat{F}_n(a) \to F_0(a)$ almost surely for each $a \in \mathcal{A}$.*

Proof. We use the following inequality given in Ayer *et al.* (1955, p. 644):

$$\sum_{a \in \mathcal{A}} \{\hat{F}_n(a) - F_0(a)\}^2 N_n(a) \leq \sum_{a \in \mathcal{A}} \{\tilde{F}_n(a) - F_0(a)\}^2 N_n(a).$$

We get

$$\sum_{a \in \mathcal{A}} \{\hat{F}_n(a) - F_0(a)\}^2 N_n(a) \leq \sum_{a \in \mathcal{A}} |N_n(a) - g(a)| + \sum_{a \in \mathcal{A}} \{\tilde{F}_n(a) - F_0(a)\}^2 g(a).$$

It follows from the SLLN that for each $a \in \mathcal{A}$, $N_n(a) \rightarrow g(a)$ and $\tilde{F}_n(a) \rightarrow F_0(a)$ almost surely. Thus Scheffé's theorem (see Billingsley 1968, p. 224) implies

$$\sum_{a \in \mathcal{A}} |N_n(a) - g(a)| \rightarrow 0 \qquad \text{almost surely}$$

and the Lebesgue dominated-convergence theorem implies

$$\sum_{a \in \mathcal{A}} \{\tilde{F}_n(a) - F_0(a)\}^2 g(a) \rightarrow 0 \qquad \text{almost surely.}$$

It follows that $\sum_{a \in \mathcal{A}} \{\hat{F}_n(a) - F_0(a)\}^2 N_n(a) \rightarrow 0$ almost surely. This yields the desired result, as $N_n(a)$ is eventually positive with probability 1 for each $a \in \mathcal{A}$. $\square$

The above result was already observed by Ayer et al. (1955) in the case when $\mathcal{A}$ is finite. In this case one can even conclude that the GMLE is uniformly strongly consistent on $\mathcal{A}$, i.e., $\sup_{a \in \mathcal{A}} |\hat{F}_n(a) - F_0(a)| \rightarrow 0$ almost surely. For countably infinite $\mathcal{A}$, however, additional assumptions are required to conclude this, as demonstrated by the following example.

Example 2.2. Suppose $\mathcal{A} = \{y_i : y_i = 1 - 1/i, \; i \geq 1\}$ and $G(y) = y$ for $y \in \mathcal{A}$. Then the GMLE will not be uniformly strongly consistent on $\mathcal{A}$ if $0 < F(1-) < 1$.

Proof. Let $\Omega_n = \bigcup_{i=1}^n \bigcap_{j \neq i} \{X_i \leq Y_i, \; Y_j < Y_i\}$. Then $\Omega_n \subset \{N^+(Y_{(n)}) = 0\}$. Since $\hat{F}_n(Y_{(n)}) = 1$ on the event $\{N^+(Y_{(n)}) = 0\}$, as observed prior to Theorem 2.1, and since $F_0(1-) < 1$, we cannot have uniform strong convergence if $\liminf_{n \to \infty} P(\Omega_n) > 0$. But

$$P(\Omega_n) = nP\left(\bigcap_{j=2}^n \{X_1 \leq Y_1, \; Y_j < Y_1\}\right) \geq n F_0(y_n)\{G(y_n-)\}^{n-1} P(Y_1 \geq y_n)$$

so that by the choice of $\mathcal{A}$ and G

$$\liminf_{n \to \infty} P(\Omega_n) \geq \liminf_{n \to \infty} F_0(y_n)\left(1 - \frac{1}{n-1}\right)^{n-1} = \frac{F(1-)}{e} > 0.$$

Consequently, the GMLE is not uniformly consistent on $\mathcal{A}$. $\square$

We now address the uniform strong consistency.

COROLLARY 2.3. *Suppose the set $\mathcal{A}$ is closed. Assume that $F_0(a-) = F_0(a)$ for each $a \in \mathcal{A}$ for which there is a strictly increasing sequence of points $\{a_i\}_{i \geq 1}$ in $\mathcal{A}$ such that $a_i \uparrow a$. Then the GMLE is uniformly strongly consistent on $\mathcal{A}$.*

Proof. Let m be a positive integer. Let $\mathcal{A}_i = \{a \in \mathcal{A} : x_{i-1} \leq a < x_i\}$, $i = 1, \ldots, m$, where $x_0 = -\infty$, $x_m = \infty$ and $x_i = \inf\{x : F_0(x) \geq i/m\}$, $i = 1, \ldots, m-1$. Let $a \in \mathcal{A}$. Then $a \in \mathcal{A}_i$ for some $i = 1, \ldots, m$. Since $\mathcal{A}$ is a closed set, $a_i = \inf \mathcal{A}_i$ and $b_i = \sup \mathcal{A}_i$ belong to $\mathcal{A}$. Using the monotonicity of $\hat{F}_n$ and F_0, we find that

$$|\hat{F}_n(a) - F_0(a)| \leq \max\{|\hat{F}_n(b_i) - F_0(b_i)|, |\hat{F}_n(a_i) - F_0(a_i)|\} + F_0(b_i) - F_0(a_i).$$

If $b_i < x_i$, then $F_0(b_i) - F_0(a_i) < 1/m$. If $b_i = x_i$, then $F_0(x_i) = F_0(x_i-) = i/m$ and $F_0(b_i) - F_0(a_i) \leq 1/m$. This shows that $\limsup_{n \to \infty} \sup_{a \in \mathcal{A}} |\hat{F}_n(a) - F_0(a)| \leq 1/m$ on

the event $\Omega_* = \bigcap_{a \in \mathcal{A}} \{ \lim_{n \to \infty} \hat{F}_n(a) = F_0(a) \}$. Since m is arbitrary and $P(\Omega_*) = 1$ by Theorem 2.1, we obtain the desired result. $\square$

In the next corollary, the set $\mathcal{A}$ need not be closed.

COROLLARY 2.4. *Assume that* $\mathcal{A} = \{a_i\}_{i \geq 1}$, *where* $a_i < a_{i+1}$ *for all* i. *Let* $\tau = \sup_i a_i$. *If* $F_0(\tau-) = 1$, *then the GMLE is uniformly strongly consistent on* $\mathcal{A}$.

Proof. Let m be a positive integer. Since

$$\sup_{a \in \mathcal{A}} | \hat{F}_n(a) - F_0(a)| \leq \max_{1 \leq i \leq m} | \hat{F}_n(a_i) - F_0(a_i)| + 1 - F_0(a_m),$$

it follows from Theorem 2.1 that $\limsup_{n \to \infty} \sup_{a \in \mathcal{A}} | \hat{F}_n(a) - F_0(a)| \leq 1 - F(a_m)$. The desired result follows, as m is arbitrary and $F_0(\tau-) = 1$. $\square$

We call a number x a *point of increase of* F_0 if either $F_0(x) < F_0(y)$ for all $y > x$ or $F_0(y) < F_0(x)$ for all $y < x$. Note that, for each α in the interval $(0, 1)$, the left quantile $F_0^{-1}(\alpha) = \inf\{ y : F(y) \geq a \}$ is a point of increase of F_0.

COROLLARY 2.5. *Suppose that* F_0 *is continuous and the closure of* $\mathcal{A}$ *contains the set* S *of all points of increase of* F_0. *Then the GMLE is uniformly strongly consistent, i.e.,* $\sup_{x \in \mathbb{R}} | \hat{F}_n(x) - F_0(x)| \to 0$ *almost surely.*

Proof. Let $F_1, F_2, \ldots$ be subdistribution functions such that $F_n(a) \to F_0(a)$ for all $a \in \mathcal{A}$. Let m be a positive integer. Since F_0 is continuous, there are points $x_1 < \cdots < x_m$ in S such that $F_0(x_i) = i/(m+1)$. The continuity of F_0 and the fact that the closure of $\mathcal{A}$ contains S imply that there are points $a_1 < \cdots < a_m$ in $\mathcal{A}$ such that $|F_0(a_i) - F_0(x_i)| \leq 1/m^2$. Using this and the monotonicity of F_0 and F_n we derive that

$$| F_n(x) - F_0(x)| \leq \max_{1 \leq i \leq m} | F_n(a_i) - F_0(a_i)| + \frac{3}{m}, \qquad x \in \mathbb{R}.$$

This shows that F_n converges to F_0 uniformly.

By the above, the events $\bigcap_{a \in \mathcal{A}} \{ \hat{F}_n(a) \to F_0(a) \}$ and $\{ \sup_{x \in \mathbb{R}} | \hat{F}_n(x) - F_0(x)| \to 0 \}$ are identical and thus have probability 1 by Theorem 2.1. $\square$

3. THE ASYMPTOTIC NORMALITY OF THE GMLE

We shall now discuss asymptotic normality and efficiency of $\hat{F}_n(x)$ for regular points x as defined next. Let $\mathcal{A}_* = \mathcal{A} \cup \{-\infty, \infty\}$. For $x \in \mathbb{R}$, set

$$x_- := \sup\{a \in \mathcal{A}_* : a < x\} \quad \text{and} \quad x_+ := \inf\{a \in \mathcal{A}_* : a > x\}.$$

We say x is a *regular point* if x belongs to $\mathcal{A}$, x_- and x_+ belong to $\mathcal{A}_*$, $x_- < x < x_+$ and $F_0(x_-) < F_0(x) < F_0(x_+)$. It is worth mentioning that there may be infinitely many regular points. For example, if F_0 is strictly increasing and $\mathcal{A}$ is the set of all positive integers, then every positive integer is a regular point. The conditions imposed on regular points are somewhat similar to the assumption that F_0 and G have positive and continuous derivatives needed in the asymptotic distribution result of the GMLE (see Groeneboom and Wellner 1992). However, their convergence rate is $n^{\frac{1}{3}}$, while we shall show that the convergence rate is $n^{\frac{1}{2}}$ under our assumptions.

Given a regular point x, $\hat{F}_n(x)$ may or may not be the same as $\bar{F}_n(x)$, as shown by the following example. Suppose that F is the exponential distribution function and $\mathcal{A} =$

$\{1, 2\}$. Then both 1 and 2 are regular points. If a sample of size 3 consists of observations $\{(1, 0), (1, 1), (2, 1)\}$, then $(\bar{F}(1), \bar{F}(2)) = (\frac{1}{2}, 1)$, which is the same as $(\hat{F}(1), \hat{F}(2))$. On the other hand, if a sample of size 3 consists of observations $\{(1, 0), (1, 1), (2, 0)\}$, then $(\bar{F}(1), \bar{F}(2)) = (\frac{1}{2}, 0)$, which is not the same as $(\hat{F}(1), \hat{F}(2)) = (\frac{1}{3}, \frac{1}{3})$. However, the following lemma shows that the two estimators differ only on a set whose probability tends to zero.

LEMMA 3.1. *Suppose x is a regular point. Then $P(\bar{F}_n(x) = \hat{F}_n(x)) \to 1$.*

Proof. Assume first that x_- and x_+ belong to $\mathcal{A}$. Let $B_n = \{\hat{F}_n(x_-) < \hat{F}_n(x) < \hat{F}_n(x_+)\}$ and $C_n = \{N_n(x_-) > 0, N_n(x) > 0, N_n(x_+) > 0\}$. It follows from Theorem 2.1 and $F_0(x_-) < F_0(x) < F_0(x_+)$ that $P(B_n) \to 1$, and from the SLLN that $P(C_n) \to 1$. In view of Theorem 2.1 in Ayer *et al.* (1955), we have, on the event $B_n \cap C_n$,

$$\hat{F}_n(x_-) < \bar{F}_n(x) \le \hat{F}_n(x) \le \bar{F}_n(x) < \hat{F}_n(x_+).$$

That is, $\hat{F}_n(x) = \bar{F}_n(x)$. Thus the desired result follows, as $P(B_n \cap C_n) \to 1$. This proves the claim when x_- and x_+ belong to $\mathcal{A}$.

If $x_+ \notin \mathcal{A}$ and $x_- \in \mathcal{A}$, then $x_+ = +\infty$, since x is a regular point. Let

$$B_n^+ = \{\hat{F}_n(x_-) < \hat{F}_n(x) < 1\} \quad \text{and} \quad C_n^+ = \{N_n(x_-) > 0, \ N_n(x) > 0\}.$$

It follows from Theorem 2.1 and $F_0(x_-) < F_0(x) < 1$ that $P(B_n^+) \to 1$, and from the SLLN that $P(C_n^+) \to 1$. In view of Theorem 2.1 in Ayer *et al.* (1955), we have, on the event $B_n^+ \cap C_n^+$, that $\hat{F}_n(x_-) < \bar{F}_n(x) \le \hat{F}_n(x) \le \bar{F}_n(x)$. That is, $\hat{F}_n(x) = \bar{F}_n(x)$. Thus the desired result follows, as $P(B_n^+ \cap C_n^+) \to 1$. This proves the claim when x_- but not x_+ belongs to $\mathcal{A}$.

The proof when x_+ but not x_- belongs to $\mathcal{A}$ is similar and will be omitted. □

The above result shows that $\hat{F}_n(x)$ has the same asymptotic properties as $\bar{F}_n(x)$. Thus the following result is immediate.

THEOREM 3.2. *Let x be a regular point. Then*

$$\hat{F}_n(x) - F_0(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{I[Y_j = x]}{g(x)} \{\Delta_j - F_0(x)\} + o_p(n^{-\frac{1}{2}}).$$

Consequently, $n^{\frac{1}{2}}\{\hat{F}_n(x) - F_0(x)\}$ is asymptotically normal with mean 0 and variance $F_0(x)\{1 - F_0(x)\}/g(x)$. This asymptotic variance can be consistently estimated by $\hat{F}_n(x)\{1 - \hat{F}_n(x)\}/N_n(x)$. Also, if $x_1 < \cdots < x_m$ are regular points, then $n^{\frac{1}{2}}\{\hat{F}_n(x_1) - F_0(x_1), \ldots, \hat{F}_n(x_m) - F_0(x_m)\}$ is asymptotically normal with mean vector 0 and diagonal covariance matrix.

Let us now address efficiency considerations. For this fix a regular point x. It follows from the above theorem that $\hat{F}_n(x)$ has influence function ψ given by

$$\psi(\Delta, Y) = \frac{I[Y = x]}{g(x)} \{\Delta - F_0(x)\}.$$

We shall now show that ψ is the efficient influence function for estimating $F_0(x)$. This will show that $\hat{F}_n(x)$ is a least-dispersed regular estimator of $F_0(x)$. The reader unfamiliar with these concepts should consult the monograph by Bickel *et al.* (1993). Let $\mathcal{H}$ be

the set of all measurable functions such that $\int h \, dF_0 = 0$ and $\int h^2 \, dF_0 < \infty$. For $h \in \mathcal{H}$ define a sequence $F_{n,h}$ of distribution functions by

$$F_{n,h}(t) = \int_{(-\infty,t]} (1 + n^{-\frac{1}{2}} h_n) \, dF_0, \qquad t \in \mathbb{R},$$

where $h_n = hI[2|h| \leq n^{\frac{1}{2}}] - \int hI[2|h| \leq n^{\frac{1}{2}}] \, dF_0$. Then

$$n^{\frac{1}{2}}\{ F_{n,h}(x) - F_0(x) \} \to H(x) = \int_{(-\infty,x]} h \, dF_0.$$

The tangent (or score function) τ_h associated with the perturbed distribution $F_{n,h}$ is given by

$$\tau_h(\Delta, Y) = H(Y)\left(\frac{\Delta}{F_0(Y)} - \frac{1-\Delta}{1-F_0(Y)} \right) = \frac{H(Y)\{\Delta - F_0(Y)\}}{F_0(Y)\{1 - F_0(Y)\}}.$$

Finally, it is easy to check that $\mathcal{E}\{\psi(\Delta, Y)\tau_h(\Delta, Y)\} = H(x)$. Since this holds for all $h \in \mathcal{H}$ and since ψ is a tangent, i.e., $\psi = \tau_h$ for some $h \in \mathcal{H}$ with $H(Y) = I[Y = x]$ $F_0(x)\{1 - F_0(x)\}/g(x)$, we obtain that ψ is the efficient influence function if G is known. However, ψ is also the efficient influence function if G is unknown, as the tangents for G are orthogonal to the tangents $\{\tau_h : h \in \mathcal{H}\}$ for F_0.

4. CONCLUDING REMARKS

The main results of our paper are given in Theorems 2.1 and 3.2. Theorem 2.1 gives the strong consistency at each point in $\mathcal{A}$, while Theorem 3.2 obtains asymptotic normality at regular points. Thus $\hat{F}_n(x)$ is both strongly consistent and asymptotically normally distributed at each regular point x. Typically, consistency fails to hold for points of increase that are not in the closure of $\mathcal{A}$. Also, the asymptotic normality result may not hold for nonregular points, as the following example shows.

Example 4.1. Assume that $\mathcal{A}$ consists of just four points, namely $a_1 < a_2 < a_3 < a_4$, and that $0 < F(a_1) < F(a_2) = F(a_3) < F(a_4) < 1$. On the event $A_n = \{\hat{F}_n(a_1) \leq \tilde{F}_n(a_2) \leq \tilde{F}_n(a_3) \leq \tilde{F}_n(a_4)\}$ we have $\hat{F}_n(a_i) = \tilde{F}_n(a_i)$, $i = 1,\ldots,4$, and on the event $B_n = \{\tilde{F}_n(a_1) \leq \tilde{F}_n(a_2) \leq \tilde{F}_n(a_4), \tilde{F}_n(a_1) \leq \tilde{F}_n(a_3) \leq \tilde{F}_n(a_4), \tilde{F}_n(a_2) > \tilde{F}_n(a_3)\}$ we have $\hat{F}_n(a_i) = \tilde{F}_n(a_i)$ for $i = 1, 4$ and $\hat{F}_n(a_2) = \hat{F}_n(a_3) = \bar{F}_n$, where

$$\bar{F}_n = \frac{N_n^-(a_2) + N_n^-(a_3)}{N_n(a_2) + N_n(a_3)}.$$

It follows from the SLLN that $P(A_n \cup B_n) \to 1$. This shows that the asymptotic distribution of $\sqrt{n}(\hat{F}_n(a_2) - F_0(a_2), \hat{F}_n(a_3) - F_0(a_3))^{\mathsf{T}}$ is the same as that of $\sqrt{n}(\tilde{F}_n^*(a_2) - F_0(a_2), \tilde{F}_n^*(a_3) - F_0(a_3))^{\mathsf{T}}$, where $(\tilde{F}_n^*(a_2), \tilde{F}_n^*(a_3)) = (\tilde{F}_n(a_2), \tilde{F}_n(a_3))$ if $\tilde{F}_n(a_2) \leq \tilde{F}_n(a_3)$, and $\tilde{F}_n^*(a_2) = \tilde{F}_n^*(a_3) = \bar{F}_n$ if $\tilde{F}_n(a_2) > \tilde{F}_n(a_3)$. An application of Slutsky's theorem yields that the asymptotic distribution of $\sqrt{n}(\tilde{F}_n^*(a_2) - F_0(a_2), \tilde{F}_n^*(a_3) - F_0(a_3))^{\mathsf{T}}$ is the distribution of the bivariate random vector Z^* defined by

$$Z^* = \begin{pmatrix} Z_1^* \\ Z_2^* \end{pmatrix} = \begin{pmatrix} Z_2 \\ Z_3 \end{pmatrix} I[Z_2 \leq Z_3] + \frac{g(a_2)Z_2 + g(a_3)Z_3}{g(a_2) + g(a_3)} \begin{pmatrix} 1 \\ 1 \end{pmatrix} I[Z_2 > Z_3],$$

where Z_2 and Z_3 are independent normal random variables with zero means and variances $F(a_2)\{1 - F(a_2)\}/g(a_2)$ and $F(a_3)\{1 - F(a_3)\}/g(a_3)$, respectively. One can check that the distributions of Z_1^* and Z_2^* are not normal.

The corollaries in Section 2 address uniform strong consistency under different sets of assumptions. Corollary 2.3 implies that the GMLE is uniformly strongly consistent on $\mathcal{A}$ if F is continuous and $\mathcal{A}$ is closed. Corollary 2.4 gives uniform consistency on $\mathcal{A}$ if this set is generated by an increasing sequence. If F is increasing and $\mathcal{A} \subset \{x \in \mathbb{R} : 0 < F(x) < 1\}$, then the assumptions of Corollary 2.4 imply that each point in $\mathcal{A}$ is regular and thus, in view of Theorem 3.2, the asymptotic normality at each point in $\mathcal{A}$.

Corollary 2.5 is of interest from a theoretical point of view in that it provides conditions that guarantee the uniform strong consistency on the entire line. From a practical point of view the imposed conditions are rather unrealistic. For example, if F is the uniform distribution on $[0,1]$, then $\mathcal{A}$ has to contain a dense subset of $[0,1]$. But distributions G with this property are rarely encountered in practice. Note also that the assumptions of Corollary 2.5 rule out the existence of regular points, so that we cannot conclude the asymptotic normality from Theorem 3.2.

It is an open question whether the parametric convergence rate holds at each point in $\mathcal{A}$. Since one can show that $\tilde{F}_n$ has parametric convergence rate at each point in $\mathcal{A}$, we conjecture that the GMLE has the same property although the limit might not be normal as Example 4.1 shows.

Groeneboom and Wellner (1992) showed that the GMLE is uniformly strongly consistent if F_0 and G are continuous and $P_{F_0} \ll P_G$. The latter means that the probability measure P_{F_0} induced by F_0 is absolutely continuous with respect to the probability measure P_G induced by G. In view of our Corollary 2.5, we expect the uniform strong consistency also if F_0 is continuous and if G is a mixture of a continuous distribution function and a discrete distribution function which satisfies the assumptions in Corollary 2.5.

Groeneboom and Wellner (1992) showed that under the additional assumption that F_0 and G have positive derivatives at a point t_0, the convergence rate of $\hat{F}_n(t_0)$ is $n^{\frac{1}{3}}$. It is an open question whether the rate $n^{\frac{1}{3}}$ is still valid without this additional assumption.

Our parametric convergence rate $n^{\frac{1}{2}}$ in Theorem 3.2 is in constrast to the nonparametric convergence rate $n^{\frac{1}{3}}$ under their assumptions. Our Theorem 3.2 is trivially true under the assumption that both X and Y take on the same finitely many values. In this case, the problem reduces to the estimation of the parameters of a multinomial distribution function, which is a parametric problem giving the usual $n^{\frac{1}{2}}$ convergence rate. This simple fact was noticed without proof by Peto (1973) and Turnbull (1976) as they both conjectured (incorrectly) that the GMLE has a convergence rate $n^{\frac{1}{2}}$ in general. We have established the parametric convergence rate of the GMLE for the first time under the assumption that X and Y may take infinitely many values.

REFERENCES

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.*, 26, 641–647.

Becker, N.G., and Melbye, M. (1991). Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity. *Austral. J. Statist.*, 33, 125–133.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins Univ. Press, Baltimore.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42, 845–854.

Groeneboom, P., and Wellner, J.A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser Verlag, Basel.

Keiding, M. (1991). Age-specific incidence and prevalence: A statistical perspective. (With discussion.) *J. Roy. Statist. Soc. Ser. A*, 154, 371–412.

Peto, R. (1973). Experimental survival curve for interval-censored data. *Appl. Statist.*, 22, 86–91.

Turnbull, B.W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser. B*, 38, 290–295.

Wang, Z., and Gardiner, J.C. (1996). A class of estimators of the survival function from interval-censored data. *Ann. Statist.*, 24, 647–658.

Department of Mathematics
State University of New York
Binghamton, New York
U.S.A. 13902

email: qyu@math.binghamton.edu

Department of Mathematics
University of New Orleans
New Orleans, Louisiana
U.S.A. 70148

Strang Cancer Prevention Center
Cornell University Medical School
New York, New York
U.S.A. 10021

ASYMPTOTIC VARIANCE OF THE GMLE OF A SURVIVAL FUNCTION WITH INTERVAL-CENSORED DATA

By QIQING YU*
State University of New York at Binghamton, New York
LINXIONG LI**
University of New Orleans, Los Angeles
and
GEORGE Y.C. WONG***
Strang Cancer Preventive Center, New York

SUMMARY. Interval-censored data are generated by a random survival time X and a random censoring interval. We either observe the exact survival time or only know the survival time lies within the censoring interval. Turnbull (1976) proposes a self-consistent algorithm for obtaining the generalized maximum likelihood estimator (GMLE) of a survival function with interval-censored data. Yu, Li and Wong (1996) prove the strong consistency of the GMLE. In this paper, we establish the asymptotic normality of the GMLE and self-consistent estimators (SCE) and present a consistent estimator of the asymptotic variance of the GMLE and SCEs with interval-censored data.

1. Introduction

We consider the nonparametric estimation of distribution function F of a survival time X with incomplete observations due to interval censoring. Interval-censored (IC) data are bivariate observations (L_i, R_i), $i = 1, ..., n$, where $L_i \le R_i$. If $L_i = R_i$, then a survival time $X_i = L_i = R_i$ is observed and we say it is an *exact observation*; if $L_i < R_i$, then X_i is censored and a censoring interval $(L_i, R_i]$ is observed instead.

SYMPTOTIC VARIANCE OF THE GMLE OF A SURVIVAL FUNCTION WITH INTERVAL-CENSORED DATA

By QIQING YU*

State University of New York at Binghamton, New York

LINXIONG LI**

University of New Orleans, Los Angeles

and

GEORGE Y.C. WONG***

Strang Cancer Preventive Center, New York

SUMMARY. Interval-censored data are generated by a random survival time X and a ... om censoring interval. We either observe the exact survival time or only know the survival ... lies within the censoring interval. Turnbull (1976) proposes a self-consistent algorithm for ... ning the generalized maximum likelihood estimator (GMLE) of a survival function with ... val-censored data. Yu, Li and Wong (1996) prove the strong consistency of the GMLE. In ... paper, we establish the asymptotic normality of the GMLE and self-consistent estimators ...) and present a consistent estimator of the asymptotic variance of the GMLE and SCEs ... interval-censored data.

1. Introduction

We consider the nonparametric estimation of distribution function F of a ... ival time X with incomplete observations due to interval censoring. Interval-... ored (IC) data are bivariate observations (L_i, R_i), $i = 1, ..., n$, where $L_i \le$ ... If $L_i = R_i$, then a survival time $X_i = L_i = R_i$ is observed and we say it is ... *exact observation*; if $L_i < R_i$, then X_i is censored and a censoring interval ... $R_i]$ is observed instead.

Recent studies of interval censoring have focussed on case 2 interval-censored (IC) data, which involve a time-to-event variable X whose value is never observed but is known to lie in the time interval between two consecutive inspections times U and V. Case 2 interval censoring arises naturally in a longitudinal follow-up study in which the event of interest cannot be easily observed (for instance, cancer recurrence, elevation of levels of a biomaker without any noticeble symptoms).

In this paper, we consider IC data which consist of both case 2 IC data and exact observations. We call such data mixed IC data. An example of such data from the Framingham Heart Study was presented by Odell et al. (1992).

To formulate a model for such data, let (Y, Z) be a random censoring vector having distribution function $G(y, z)$, and let $T \ge 0$ be a random variable having distribution function $G_T(t)$. Assume that $0 \le Y < Z \le \tau_T$, with probability one, where $\tau_T = \sup\{t; P(\min(X, T) \le t) < 1\}$. We say a data point is from a case 2 interval censorship (C2) model (Groeneboom and Wellner, 1992) if the corresponding observation is $(Y, Z, 1(X \le Y), 1(X \le Z))$, where $1(\cdot)$ is the indicator function; and a data point is from a right censorship (RC) model if the corresponding observation is $(\min(X, T), 1(X \le T))$. Thus an observation obtained from the C2 model is always censored and that from the RC model is either right-censored or exact. IC data can be viewed as a mixture of data from a C2 model and a RC model. We introduce a random variable, D, to distinguish failure times coming from the two models:

$$D = \begin{cases} 1 & \text{if the observation is from the RC model,} \\ 0 & \text{if the observation is from the C2 model.} \end{cases}$$

Let $P\{D = 1\} = \pi$. We assume $0 < \pi \le 1$ and D, X, T and (Y, Z) are independent. Formally, the observable random interval $\{L, R\}$ can be expressed as follows.

$$\{L, R\} = \begin{cases} (0, Y] & \text{if } D = 0 \text{ and } X \le Y, \\ (Y, Z] & \text{if } D = 0 \text{ and } Y < X \le Z, \\ (Z, \infty) & \text{if } D = 0 \text{ and } X > Z, \\ (T, \infty) & \text{if } D = 1 \text{ and } X > T, \\ [X, X] & \text{if } D = 1 \text{ and } X \le T. \end{cases} \quad \dots (1.1)$$

We call such a model an IC model.

Under the RC model, the product limit estimator (Kaplan and Meier, 1958) of F is the generalized maximum likelihood estimator (GMLE), which maximizes the generalized likelihood function, and has been studied by many authors. Under the C2 model and under certain smoothness conditions, Groeneboom and Wellner (1992) show that the GMLE of F is strongly consistent and conjecture that the convergence rate of the GMLE is $(n \ln n)^{1/3}$. With IC data, Peto (1973) proposes the GMLE $\hat{F}$ of F. Turnbull (1976) proposes a self-consistent algorithm for obtaining the GMLE of F. It is known that the GMLE is a self-consistent estimator (SCE) (see, e.g., Gu and Zhang, 1993). Yu, Li and Wong (1996)

ablish the strong consistency of the GMLE and SCEs under the following
umption.

ASSUMPTION 1. F is arbitrary and (Y, Z, T) takes on finitely many values.

However, the asymptotic distribution of the GMLE based on IC data has
been discussed. Due to lack of consistent estimators of the asymptotic
iance of the GMLE, the application of the GMLE of $\hat{F}$ with IC data has
n limited. A common practice in medical research with IC data is to treat
right endpoint r_i as an exact observation if r_i is not ∞ (see, e.g., Samuelsen
Kongerud, 1994). Then the data set is treated as a right-censored data set
thus standard statistical tools can be used.

In Section 2, we introduce notations and necessary background. In Section
we prove the asymptotic normality of the GMLE and SCEs and present a
nsistent estimator of the asymptotic variance. The proof of a technical lemma
Section 3 is provided in the appendix.

2. Notations and Definitions

We shall establish our results in two steps. First obtain the asymptotic
ults under the following stronger assumption.

ASSUMPTION 2. (L, R) takes on finitely many distinct values $(l_1, r_1), ..., (l_g, r_g)$,
th probabilities $p_1, ..., p_g$, where $p_i > 0$ and $\sum_{i=1}^g p_i = 1$.

Then, at the second step we extend the result to the case where a weaker
sumption, Assumption 1, is assumed.

Suppose that the IC data consist of N_1 pairs of $\{l_1, r_1\}$, N_2 pairs of $\{l_2, r_2\}$, ...,
d N_g pairs of $\{l_g, r_g\}$, where $N_i \geq 0$ and $N_1 + ... + N_g = n$. Let $I_1, ..., I_n$
note the observed intervals. Define $innermost\ intervals$ $A_j, j = 1, 2, ..., m$
duced by $I_1, ..., I_n$ to be all the disjoint intervals which are non-empty in-
rsections of these I_i's such that for all possible i and j, $A_j \cap I_i = \emptyset$ or A_j.
t the endpoints of the innermost intervals be a_j and $b_j, j = 1, ..., m$, where
$\leq b_1 \leq a_2 \leq b_2 \leq ... \leq a_m \leq b_m$. Let $\delta_{ij} = 1(A_j \subset I_i)$. The following
ample illustrates the procedure of finding innermost intervals.

EXAMPLE 2.1. Suppose that there are five observed intervals $(1, 4]$, $[2, 2]$,
6], [5, 5], and (1, 6]. Then there are two exact observations, $I_1 = [2, 2]$ and
$= [5, 5]$, and three censored intervals, $I_3 = (1, 4]$, $I_4 = (2, 6]$ and $I_5 = (1, 6]$.
rthermore, there are three innermost intervals, $A_1 = [2, 2]$, $A_2 = [5, 5]$, and
3 $= (2, 4]$.

Peto (1973) shows that the GMLE of F assigns weight, say $s_1, ..., s_m$, to the
rresponding innermost intervals $A_1, ..., A_m$ only. The generalized likelihood

function L^* (Kiefer and Wolfowitz, 1956) can be simplified as

$$L^* = L^*(s_1, ..., s_{m-1}) = \prod_{i=1}^n [\sum_j \delta_{ij} s_j], \qquad ...(2.1)$$

where $s^t = (s_1, ..., s_{m-1}) \in D_s$ is the transpose of s, $D_s = \{s;\ s_i \geq 0, s_1 + ... + s_{m-1} \leq 1\}$ and $s_m = 1 - s_1 - ... - s_{m-1}$. Turnbull (1976) proposes a self-consistent algorithm for obtaining the GMLE via an iterative procedure as follows. At step 1, let $s_j^{(1)} = 1/m$ for $j = 1, ..., m$. At step h,

$$s_j^{(h)} = \sum_{i=1}^n \frac{1}{n} \frac{\delta_{ij} s_j^{(h-1)}}{\sum_{k=1}^m \delta_{ik} s_k^{(h-1)}}, \quad j = 1, ..., m.$$

He shows that, as $h \to \infty$, $s_j^{(h)}$ converges to the GMLE, $\hat{s}_j$, which maximizes L^* and satisfies the system of self-consistent equations

$$\hat{s}_j = \sum_{i=1}^n \frac{1}{n} \frac{\delta_{ij} \hat{s}_j}{\sum_{k=1}^m \delta_{ik} \hat{s}_k}, \quad j = 1, ..., m. \qquad ...(2.2)$$

A solution $s = \hat{s}$ to (2.2) is called a self-consistent estimator of s if $\hat{s} \in D_s$.
An estimate $\hat{F}(t)$ of $F(t)$ can be uniquely defined for $t \in [b_i, a_{i+1}]$ by $\hat{F}(b_i) = \hat{F}(a_{i+1}-) = \hat{s}_1 + ... + \hat{s}_i$, but is not uniquely defined for t being in an open innermost interval (Peto, 1973; Turnbull, 1976). To avoid this ambiguity we define

$$\hat{F}(t) = \begin{cases} \hat{s}_1 + ... + \hat{s}_i & \text{if } t \in (b_i, a_{i+1}], \\ \hat{s}_1 + ... + \hat{s}_{i-1} + \hat{s}_i \frac{t - a_i}{b_i - a_i} & \text{if } t \in (a_i, b_i], a_i > 0 \text{ and } b_i \neq \infty, \\ \hat{s}_1 & \text{if } 0 = a_1 \leq t \leq b_1, \\ 1 - \hat{s}_m & \text{if } t \geq a_m \text{ and } a_m < b_m = \infty, \end{cases} \qquad ...(2.3)$$

where $(a, b]$ is an empty set if $a = b$. Under the RC model, the definition of $\hat{F}$ given by (2.3) reduces to the PLE (Gill, 1983) for $t \geq a_m$ when $b_m = \infty$, and possesses the uniform strong consistency. If we define $\hat{F}(t) = 0$ for $t \in [0, b_1]$ when $a_1 = 0$, or $\hat{F}(t) = 1$ for $t \in [a_m, \infty)$ when $b_m = \infty$, the uniform strong consistency of $\hat{F}(t)$ does not hold (see Yu and Li, 1994 or Stute and Wang, 1993).

Let $A_j^o, j = 1, ..., m_o$ be the innermost intervals induced by the g intervals $\{l_i, r_i\}, i = 1, 2, ..., g$. To distinguish A_j^o from the innermost intervals A_j induced by the observed I_i's, we call A_j^o $population\ innermost\ intervals$. Let s_j^o be the weight assigned by the distribution function F to A_j^o, i.e., $s_j^o = P\{X \in A_j^o\}$. Note that $m_o = m$ or $A_j = A_j^o$ may not be true, since N_i may be zero for some i. However, for every i, $\lim_{n \to \infty} N_i/n = p_i > 0$ almost surely and we

vestigating the asymptotic properties of $\hat{F}$, thus, by taking large sample, we can without loss of generality (WLOG) assume that $N_i > 0$ for all , 2,...,g. As a consequence, we have $m_o = m$ and $A_j = A_j^o$. WLOG, we ssume $I_i = \{l_i, r_i\}$, $i = 1,...,g$. Then (2.3) can be changed to the following alent form,

$$\hat{s}_j = \sum_{i=1}^{g} \frac{N_i}{n} \frac{\delta_{ij}\hat{s}_j}{\sum_{k=1}^{m}\delta_{ik}\hat{s}_k}, \quad j = 1,...,m, \qquad \ldots (2.4)$$

cing N_i/n in (2.4) by p_i, we obtain

$$s_j = \sum_{i=1}^{g} p_i \frac{\delta_{ij}s_j}{\sum_{k=1}^{m}\delta_{ik}s_k}, \quad j = 1,...,m. \qquad \ldots (2.5)$$

following theorem is proved by Yu, Li and Wong (1996).

THEOREM 1. *Under the IC model and Assumption 2,* $s^o = (s_1^o,...,s_m^o)$ *is nique solution of the equations (2.5), and for all SCE $\hat{s}$,* $\lim_{n\to\infty}\hat{s}_i = s_i^o$ *st surely,* $i = 1,2,...,m.$

3. Main Result

n this section, we develop the asymptotic normality of the GMLE and SCEs d $\hat{F}(t)$ under the IC model. We first consider Assumption 2. By the end of section, we will relax it to Assumption 1.

3.1. *Results under Assumption 2.* Note that each value $x_i \in [\eta, \tau_r]$ of X stitutes a (population) innermost interval, where $\eta = \sup\{x : F(x) = 0\}$. pose that $A_1^o < \cdots < A_m^o$ are the population innermost intervals. For the r development, we need to assume

$$s_i^o > 0 \text{ for } i = 1,...,m, \text{ and } \sum_{i=1}^{m} s_i^o = 1 \qquad \ldots (3.1)$$

us we assume that there exist at least $m-1$ points $x_1, x_2,..., x_{m-1} \in [\eta, \tau_r]$ h that $x_1 < x_2 < \cdots < x_{m-1}$ and $A_i^o = \{x_i\}$. Under Assumption 2, assump- n (3.1) is equivalent to

ASSUMPTION 3. $P\{X \in (l,r)\} > 0$ for all $l < r$, where l and r are realizations L and R, respectively.

Let $\hat{p} = (\hat{p}_1,...,\hat{p}_{g-1})$, then $(N_1,...,N_g)$ follows a multinomial distribution $M(n; p_1,...,p_g)$. Denote $\Sigma_{\hat{p}}$ the $(g-1)\times(g-1)$ covariance matrix of $\hat{p}$, that is,

$$\Sigma_{\hat{p}} = (\alpha_{ij})_{(g-1)\times(g-1)}, \quad \alpha_{ij} = \begin{cases} p_i(1-p_i)/n & \text{if } i = j, \\ -p_i p_j/n & \text{otherwise.} \end{cases}$$

Then, it follows from Rao (1973, p.382) that

$$\Sigma_{\hat{p}}^{-1/2}(\hat{p} - p) \xrightarrow{\mathcal{D}} N(\mathbf{O}, \Im) \qquad \ldots (3.2)$$

where $\mathbf{O}$ is a $(g-1)\times 1$ zero vector and $\Im$ is a $(g-1)\times(g-1)$ identity matrix.

To find the asymptotic distribution of $\hat{s}$, we use the result (3.2) and the connection between p and s. Theorem 1 implies that s can be implicitly expressed as a function of p. For convenience we use $(s_1,...,s_m)$ instead of $(s_1^o,...,s_m^o)$ for the rest of the paper. Since $s_j > 0$ for all j, (2.5) becomes

$$1 = \sum_{i=1}^{g} p_i \frac{\delta_{ij}}{\sum_{k=1}^{m}\delta_{ik}s_k}, \quad j = 1,...,m. \qquad \ldots (3.3)$$

Taking partial derivatives $\frac{\partial}{\partial p_h}$ on both sides of the above equations yields

$$0 = \sum_{i=1}^{g}\frac{\partial p_i}{\partial p_h}\frac{\delta_{ij}}{\sum_{k=1}^{m}\delta_{ik}s_k} - \sum_{i=1}^{g}p_i\frac{\delta_{ij}\sum_{l=1}^{m}(\delta_{il}\frac{\partial s_l}{\partial p_h})}{(\sum_{k=1}^{m}\delta_{ik}s_k)^2}$$

$$= \frac{\delta_{hj}}{\sum_{k=1}^{m}\delta_{hk}s_k} - \frac{\delta_{gj}}{\sum_{k=1}^{m}\delta_{gk}s_k} - \sum_{i=1}^{g}p_i\frac{\delta_{ij}(\sum_{l=1}^{m-1}\delta_{il}\frac{\partial s_l}{\partial p_h} - \delta_{im}\frac{\partial s_l}{\partial p_h})}{(\sum_{k=1}^{m}\delta_{ik}s_k)^2}$$

$$= \frac{\delta_{hj}}{\sum_{k=1}^{m}\delta_{hk}s_k} - \frac{\delta_{gi}}{\sum_{k=1}^{m}\delta_{gk}s_k} - \sum_{i=1}^{g}p_i\frac{\delta_{ij}\sum_{l=1}^{m-1}[\delta_{il} - \delta_{im}/(m-1)]\frac{\partial s_l}{\partial p_h}}{(\sum_{k=1}^{m}\delta_{ik}s_k)^2},$$

$j = 1,...,m-1, \quad h = 1,...,g-1.$ Thus,

$$\sum_{l=1}^{m-1}\{[\sum_{i=1}^{g}p_i\frac{\delta_{ij}[\delta_{il} - \delta_{im}/(m-1)]}{(\sum_{k=1}^{m}\delta_{ik}s_k)^2}]\frac{\partial s_l}{\partial p_h}\} = \frac{\delta_{hj}}{\sum_{k=1}^{m}\delta_{hk}s_k} - \frac{\delta_{gi}}{\sum_{k=1}^{m}\delta_{gk}s_k} \qquad \ldots (3.4)$$

$= 1,...,m-1, h = 1,...,g-1$. Rewrite the system of linear equations in (3.4)

$$C\left(\frac{\partial s}{\partial p_1}, \cdots, \frac{\partial s}{\partial p_{g-1}}\right) = (w_1, \cdots, w_{g-1}),$$ $\quad$ ...(3.5)

here

$$:= (c_{jl})_{(m-1)\times(m-1)} \quad \text{with} \quad c_{jl} = \sum_{i=1}^{g} p_i \frac{\delta_{ij}[\delta_{il} - \delta_{im}/(m-1)]}{\left(\sum_{k=1}^{m}\delta_{ik}s_k\right)^2}, \quad \frac{\partial s}{\partial p_h} = \begin{pmatrix} \frac{\partial s_1}{\partial p_h} \\ \vdots \\ \frac{\partial s_{m-1}}{\partial p_h} \end{pmatrix}$$

$$'_h = \begin{pmatrix} w_{1h} \\ \vdots \\ w_{(m-1)h} \end{pmatrix} \quad \text{and} \quad w_{jh} = \frac{\delta_{hj}}{\sum_{k=1}^{m}\delta_{hk}s_k} - \frac{\delta_{gj}}{\sum_{k=1}^{m}\delta_{gk}s_k}, \quad h = 1,...,g-1, j = 1,...,m-$$

Note that C is independent of h. Rewrite

$$C = \left(\sum_{i=1}^{g} p_i \frac{\delta_{ij}(\delta_{il} - \delta_{im}/(m-1))}{\sum_{k=1}^{m}(\delta_{ik}s_k)^2}\right)_{(m-1)\times(m-1)} = UDV,$$

here

$$r = \begin{pmatrix} \delta_{11} & \cdots & \delta_{g1} \\ \vdots & \vdots & \vdots \\ \delta_{1(m-1)} & \cdots & \delta_{g(m-1)} \end{pmatrix}, \quad D = \begin{pmatrix} \frac{p_1}{\sum_{k=1}^{m}\delta_{1k}s_k)^2} & 0 & \cdots & 0 \\ 0 & \cdot & & \vdots \\ \vdots & & \cdot & 0 \\ 0 & \cdots & & \frac{p_g}{\sum_{k=1}^{m}\delta_{gk}s_k)^2} \end{pmatrix}$$

nd

$$V = \begin{pmatrix} \delta_{11} - \delta_{1m}/(m-1) & \cdots & \delta_{1(m-1)} - \delta_{1m}/(m-1) \\ \vdots & \vdots & \vdots \\ \delta_{g1} - \delta_{gm}/(m-1) & \cdots & \delta_{g(m-1)} - \delta_{gm}/(m-1) \end{pmatrix}.$$

ly relabeling the observed intervals I_i, we can assume that $I_j = A_j^o$ for $1 \le j \le i-1$. In other words, $\delta_{ij} = 1$ if $i = j$ and 0 otherwise for $1 \le i, j \le m-1$. Thus $U = (\Im, U_2)$ where $\Im$ is the $(m-1)\times(m-1)$ identity matrix and U_2 is a $m-1)\times(g-m+1)$ matrix. Therefore, the rank of U is of $m-1$. Similarly re can show that the rank of V is also $m-1$ by observing that $\delta_{im} = 0$ for $\le i \le m-1$. The rank of the matrix D is g ($\ge m$) since all $p_i > 0$. Thus, it an be shown that C is nonsingular (see Lemma A in the appendix). Moreover,

---

by letting $I_g = A_m$,

$$(w_1,...,w_{(g-1)}) = \begin{pmatrix} \frac{1}{\sum_{k=1}^{m}\delta_{1k}s_k} & \cdots & 0 & \cdots \\ & & & \\ 0 & \cdots & \frac{1}{\sum_{k=1}^{m}\delta_{(m-1)k}s_k} & \cdots \end{pmatrix}$$ $\quad$ ...(3.6)

and thus it is of rank $m-1$. It follows from (3.5) that $\frac{\partial s}{\partial p_h} = C^{-1}w_h$, $h = 1,...,g-1$. By the first order Taylor expansion, $\hat{s} - s \approx \frac{\partial s}{\partial p}(\hat{p} - p)$. The validity of the approximation can be verified by the fact that the second order partial derivatives $\frac{\partial^2 s_i}{\partial p_t\partial p_h}$ exist and are continuous. In fact, taking partial derivatives on both sides of (3.5) yields

$$U(\frac{\partial}{\partial p_t}D)V\frac{\partial s}{\partial p_h} + C\frac{\partial^2 s}{\partial p_t\partial p_h} = \frac{\partial}{\partial p_t}w_h, \quad t,h = 1,...,g-1.$$

Thus

$$\frac{\partial^2 s}{\partial p_t\partial p_h} = C^{-1}(\frac{\partial}{\partial p_t}w_h - U(\frac{\partial}{\partial p_t}D)V\frac{\partial s}{\partial p_h}), \quad t,h = 1,...,g-1,$$ $\quad$ ...(3.7)

where

$$\frac{\partial}{\partial p_t}w_h = \begin{pmatrix} \frac{\delta_{hj}\sum_{l=1}^{m}\delta_{hl}\frac{\partial s_l}{\partial p_t}}{\left(\sum_{k=1}^{m}\delta_{hk}s_k\right)^2} - \frac{\delta_{gj}\sum_{l=1}^{m}\delta_{gl}\frac{\partial s_l}{\partial p_t}}{\left(\sum_{k=1}^{m}\delta_{gk}s_k\right)^2} \end{pmatrix}_{(m-1)\times 1}$$

and $\frac{\partial}{\partial p_i}D$ is a diagonal matrix with the i-th diagonal element

$$\frac{1(i = t)}{\left(\sum_{k=1}^{m}\delta_{ik}s_k\right)^2} - 2\frac{p_i\left(\sum_{l=1}^{m}\delta_{il}\frac{\partial s_l}{\partial p_t}\right)}{\left(\sum_{k=1}^{m}\delta_{ik}s_k\right)^3}.$$

In view of (3.7), the second order partial derivatives of $s = s(p)$ exist. Thus, by (2.5), (3.6) and the well-known result of multivariate normal convergence theorem (see, for example, Rao, 1973, p.387), we have the following result.

THEOREM 2. *Under the IC model and Assumptions 2 and 3*

$$\Sigma_{\hat{s}}^{-1/2}(\hat{s} - s) \xrightarrow{\mathcal{D}} N(O, \Im), \quad as \ n \to \infty, \quad where \quad \Sigma_{\hat{s}} = (\frac{\partial s}{\partial p})\Sigma_{\hat{p}}(\frac{\partial s}{\partial p})^t,$$

is a $(m-1) \times 1$ zero vector and $\Im$ is a $(m-1) \times (m-1)$ identity matrix. Furthermore, define $e_x = (e_1, e_2, \cdots, e_{m-1})$ where $e_i = 1$ if $x \geq x_i$ and 0 otherwise. Then (2.3) and Theorem 2 imply the next theorem.

THEOREM 3. Under the assumptions of Theorem 2, we have,

$$(e_x \Sigma_{\hat{s}} e_x^t)^{-1/2}(\hat{F}(x) - F(x)) \xrightarrow{\mathcal{D}} N(0,1) \quad as \; n \to \infty,$$

for $x \in B$, where $\hat{F}(x)$ is given by (2.9) and $B = \{x : F(x) \in (0,1), \; x \leq \tau_r\}$.

Verify that $\hat{F}(x) = 0$ for $x < \tau_l$ (and thus $x \notin B$). It is obvious that then the theorem is not true. An estimator $\hat{\Sigma}_{\hat{s}}$ of $\Sigma_{\hat{s}}$ can be obtained by replacing $\hat{p}$ and $\hat{s}$ in $\Sigma_{\hat{s}}$ by $\hat{p}$ and $\hat{s}$, respectively. By the strong consistency of $\hat{p}$ and theorem 1 it is easy to see that the estimator $\hat{\Sigma}_{\hat{s}}$ is strongly consistent. Then, an immediate consequence of Theorem 2, we have the following.

THEOREM 4. Under the assumptions of Theorem 2, we have,

$$\frac{\hat{F}(x) - F(x)}{\hat{\sigma}} \xrightarrow{\mathcal{D}} N(0,1)$$

as $n \to \infty$ for $x \in B$, where $\hat{\sigma}^2 = e_x \hat{\Sigma}_{\hat{s}} e_x^t$.

3.2 Results under Assumption 1. In this subsection, we show that the results in section 3.1 are also valid under Assumptions 1 and 3. For simplicity, we only give the proof of the extension of Theorem 3.

THEOREM 5. Under the IC model and Assumptions 1 and 3,

$$(e_x \Sigma_{\hat{s}} e_x^t)^{-1/2}(\hat{F}(x) - F(x)) \xrightarrow{\mathcal{D}} N(0,1) \quad as \; n \to \infty, \quad for \; x \in B. \quad \ldots (3.8)$$

PROOF. For any arbitrary distribution function F and fixed $x_0 \in [\eta, \tau_r]$, choose finitely many points, say $\xi_1 < \cdots < \xi_{(M/2)+1}$, such that x_0, T, Y, $Z \in \{\xi_1, \ldots, \xi_{(M/2)+1}\}$ w.p.1 with $\xi_1 = \eta$, $\xi_{M/2} = \tau_r$ and $\xi_{(M/2)+1} = \infty$, where M is an even positive integer. Denote $B_{2i-1} = \{\xi_i\}$ and $B_{2i} = (\xi_i, \xi_{i+1})$, $i = 1, \ldots, (M/2)$, and denote the midpoint of the set B_j by m_j, where the midpoint of an infinite interval $(b, +\infty)$ is defined to be $b+1$. Let $B_{k_1}, \ldots, B_{k_{m^*}}$ (for some m^*) be all the distinct elements of $\{B_k : k = 1, \ldots, M\}$ such that $s_i^* = \mathrm{P}(X \in B_{k_i}) > 0$. Define a new distribution function F^* by $F^*(t) = \sum_{i: m_{k_i} \leq t} s_i^*$, and define pseudo observations

$$I_j^* = \begin{cases} I_j & \text{if } I_j \text{ is non-singleton,} \\ \{m_i\} & \text{if } I_j \text{ is a singleton and } I_j \subset B_i. \end{cases}$$

Let $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_m)$ ($\hat{F}$) be an SCE of s (F) based on observations I_i, $i = 1, \ldots, n$, i.e., $\hat{s}$ satisfies (2.2). Let $A_1 < \cdots < A_m$ be the innermost intervals induced by

the original data. As assumed, n is large, thus all innermost intervals, except perhaps A_m, are singleton set due to Assumption 3. Let $A_i^* = \{m_{k_i}\}$, $i = 1, \ldots, m^* - 1$ and

$$A_{m^*}^* = \begin{cases} \{m_{k_{m^*}}\} & \text{if } \sup\{x : x \in B_{k_{m^*}}\} < +\infty, \\ B_{k_{m^*}} & \text{otherwise.} \end{cases}$$

Then the definition of $\{\xi_i\}$ implies that

$$I_i \cap A_j^* = \text{either } \emptyset \text{ or } A_j^* \text{ or } I_i. \quad \ldots (3.9)$$

WLOG, we can assume that $I_1^*, \ldots, I_n^*$ are all the distinct elements among $I_1^*, \ldots, I_n^*$. Verify that $A_1^*, \ldots, A_{m^*}^*$ are all the innermost intervals induced by $I_1^*, \ldots, I_{g^*}^*$. Let $\tilde{s}$ be such that

$$\tilde{s}_j = \sum_{h: A_h \subset B_{k_j}} \hat{s}_h, \quad j = 1, \ldots, m^*.$$

It follows from (3.9) and (3.10) that $\tilde{s}$ satisfies equations

$$\tilde{s}_j = \sum_i \frac{N_i^*}{n} \frac{\delta_{ij}^* \tilde{s}_j}{\sum_{k=1}^{m^*} \delta_{ik}^* \tilde{s}_k}, \quad j = 1, \ldots, m^*,$$

where the summation is over distinct I_i^*'s, $\delta_{ij}^* = 1(A_j^* \subset I_i^*)$ and $N_i^* = \sum_{j=1}^n 1(I_j^* = I_i^*)$. Verify that

$$N_i^* = \begin{cases} N_j & \text{if } I_i^* \text{ is a non-singleton and } I_i^* = I_j, \\ \sum_{j=1}^{N_j} N_j 1(I_j \subset B_{k_h}, I_j \text{ is a singleton}) & \text{if } I_i^* = \{m_{k_h}\}, \end{cases}$$

for $i = 1, \ldots, g^*$. Thus $\tilde{s}$ is an SCE of $s^* = (s_1^*, \ldots, s_{m^*}^*)$ based on the pseudo observations I_i^*, $i = 1, \ldots, n$. It is important to note that 1. F^* takes on finitely many ($\leq m^*$) values,

2. $s_j^* = dF^*(A_j^*) = dF(B_{k_j}) > 0$, and

3. pseudo observations $I_1^*, \ldots, I_n^*$, generated by F^*, G_T and G, are i.i.d. since original observations $I_1, \ldots, I_n$, generated by F, G_T and G, are i.i.d. Thus Theorems 1, 2 and 3 apply to $\tilde{s}$. It is readily seen from the definitions of $\{\xi_i\}$ and $\tilde{s}$ that $\hat{F}(x_0) = \sum_{j: m_{k_j} \leq x_0} \tilde{s}_j$, thus

$$(e_{x_0} \Sigma_{\hat{s}} e_{x_0}^t)^{-1/2}(\hat{F}(x_0) - F(x_0)) \xrightarrow{\mathcal{D}} N(0,1) \quad as \; n \to \infty.$$

Furthermore,

$$p_i^* = \begin{cases} p_i(1-\epsilon) & \text{if } i = 1, \dots, m-1, \\ \epsilon & \text{if } i = m, \\ p_{i-1}(1-\epsilon) & \text{if } i = m+1, \dots, g+1. \end{cases}$$

The new C^* matrix is now $m \times m$ dimensional and it can be verified that

$$C^* = \begin{pmatrix} C/(1-\epsilon) & O \\ O & \dfrac{p_{g+1}^*}{\epsilon} \end{pmatrix} \qquad \dots (A.1)$$

where p_i^* is defined in an obvious way and C^* is the matrix as before. Taking partial derivatives $\frac{\partial}{\partial p_h^*}$ on both sides of equations (3.5) after replacing C by C^*, etc. yields (letting $1 - \epsilon = q$)

$$0 = \sum_{i=1}^{g+1} \frac{\partial p_i^*}{\partial p_h^*} \frac{\delta_{ij}^*}{\sum_{k=1}^{m+1} \delta_{ik}^* s_k^*} - \sum_{i=1}^{g+1} p_i^* \frac{\delta_{ij}^* \sum_{l=1}^{m+1}(\delta_{il}^* \frac{\partial s_l^*}{\partial p_h^*})}{(\sum_{k=1}^{m+1} \delta_{ik}^* s_k^*)^2}$$

$$(\text{noting } \delta_{mj}^* = 0 \text{ if } j < m, \ \delta_{im}^* = 0 \text{ if } i \neq m \text{ and } \tfrac{\partial s_m^*}{\partial p_i^*} = 0)$$

$$= \frac{\delta_{hj}}{\sum_{k=1}^{m} \delta_{hk} q s_k} - \frac{\delta_{gj}}{\sum_{k=1}^{m} \delta_{gk} q s_k} - \sum_{i=1}^{g} q p_i \frac{\delta_{ij}(\sum_{l=1}^{m-1} \delta_{il}(\frac{\partial q s_l}{\partial q p_h} - \delta_{im} \frac{\partial q s_l}{\partial q p_h}))}{q(\sum_{k=1}^{m} \delta_{ik} s_k)^2}$$

$$= \frac{\delta_{hj}}{\sum_{k=1}^{m} \delta_{hk} q s_k} - \frac{\delta_{gj}}{\sum_{k=1}^{m} \delta_{gk} q s_k} - \sum_{i=1}^{g} p_i \frac{\delta_{ij} \sum_{l=1}^{m-1}[\delta_{il} - \delta_{im}/(m-1)]\frac{\partial s_l}{\partial p_h}}{(\sum_{k=1}^{m} \delta_{ik} s_k)^2},$$

$j = 1, \dots, m-1$, $h = 1$.

Next we rearrange the order of A_j^*'s as follows: Let $A_j^e = A_j$ for $j = 1, \dots, m$ and $A_{m+1}^e = [b, b]$; let $I_i^e = I_i$ for $i = 1, \dots, g$ and $I_{g+1}^e = [b, b]$ and define p_i^e correspondingly. The new $m \times m$ matrix C^e induced by this arrangement has the (j, l) entry

$$\lambda_{jl}^e = \sum_{i=1}^{g} p_i^e \frac{\delta_{ij}^e \delta_{il}^e}{(\sum_{k=1}^{m} \delta_{ik}^e (s_k^e))^2}, \quad j, l = 1, \dots, m,$$

$j = 1, \dots, m$. Write $C^e = U^e D^e (U^e)^t$ and it can be shown that $U^e = (\Im, U_2)$, thus C^e is nonsingular. Note that since $\delta_{ij}^e \delta_{im}^e = 0$ for $i = 1, 2, \dots, g+1$ and $j = 1, \dots, m$ in an obvious way, then it can be shown that $U^e = (\Im, U_2)$, thus C^e is nonsingular. Note that since $\delta_{ij}^e \delta_{im}^e = 0$ in an obvious way. This implies that the derivative $\frac{\partial s^e}{\partial p_i^e}$ exists and is unique. Note that singular. This implies that the derivative $\frac{\partial s^e}{\partial p_i^e}$ exists and is unique. Note that

---

; follows from (2.3) that for $x \leq \tau_r$

$$(e_x \Sigma_{\hat{S}} e_x^t)^{-1/2} (\hat{F}(x) - F(x)) \xrightarrow{D} N(0,1) \text{ as } n \to \infty. \quad \square$$

3.3 *Comments.* Some comparisons between the existing results for the C2 model and our results here are provided below.

Groeneboom and Wellner (1992) conjectured that the convergence of the GMLE is at a slower rate of $(n \log n)^{1/3}$ under the C2 model with the assumption that all the random variables are absolutely continuous. Under our assumption 2, for both C2 and IC models, the estimation of F becomes a parametric estimation problem of a finite dimensional multinomial distribution. Consequently, the GMLE should have the usual parametric $n^{1/2}$ convergence rate. Note that under the C2 model one can only estimate F at the values of Y or Z, while under the IC model we can estimate F at all values of X that are within $[\eta, \tau_r]$ because of the existence of exact observations.

Under the C2 model and assumptions 2 and 3, Yu et al. (1998) show that the GMLE of $\hat{F}(x)$ satisfies (3.8) for x at the values of Z or Y. In addition, it is worth noting that under assumptions 2 and 3, the solution to the limiting self-consistent equation (2.5) may not be unique, i.e., Theorem 1 fails to be true under the C2 model. Thus (3.8) may not be true for arbitrary SCEs under the C2 model. However, under the IC model and assumptions 2 and 3, as we have seen in this paper, all SCEs are consistent and satisfy (3.8), due to exact observations.

Appendix

LEMMA A. *The matrix C in (3.5) is nonsingular.*

PROOF. Use the notations previously defined in Section 2. By Assumption 2, WLOG, we can assume that the supports of X, T, Y and Z are bounded by some positive number b. Let X^* be a new random variable, independent of (X, T, Y, Z), with distribution function $F_*(x) = (1 - \epsilon)F(x) + \epsilon 1(x \leq b)$, where $\epsilon \in (0, 1)$.

Define $A_i^* = A_i$ for $i = 1, \dots, m-1$, where A_i are defined as before, and define $A_m^* = [b, b]$ and $A_{m+1}^* = A_m$. Let T, Y and Z be as before. Then there are $g+1$ distinct values of (L^*, R^*), where L^*, etc. are defined accordingly in an obvious way. Let the intervals induced by these pairs be

$$I_i^* = \begin{cases} I_i & \text{if } i = 1, \dots, m-1, \\ [b, b] & \text{if } i = m, \\ I_{i-1} & \text{if } i = m+1, \dots, g+1. \end{cases}$$

The new weight

$$s_j^* = \begin{cases} s_j(1-\epsilon) & \text{if } j = 1, \dots, m-1, \\ \epsilon & \text{if } j = m, \\ s_m(1-\epsilon) & \text{if } j = m+1. \end{cases}$$

$$\mathbf{s}^e=(s_1^e,\dots,s_m^e,s_{m+1}^e)=(s_1^*,\dots,s_{m+1}^*,s_m^*) \text{ and}$$

$$p_i^*=\begin{cases} p_i^e & \text{if } i=1,\dots,m-1,\\ p_{g+1}^e & \text{if } i=m,\\ p_{i-1}^e & \text{if } i=m+1,\dots,g+1,\end{cases}$$

and that the system of equations (2.5) related to s^* (note that there are $m+1$ rather than m equations) is equivalent to that related to s^e in the following sense:

Choosing any $m-1$ equations of it to derive $\frac{\partial s^*}{\partial \mathbf{p}^*}$ would be the same. C^* corresponds to choosing $i=1,\dots,m$, and C^e corresponds to choosing $i=1,\dots,m-1$ and $m+1$. Hence, even though $C^*\ne C^e$, $\frac{\partial s^e}{\partial \mathbf{p}_i^e}=(C^e)^{-1}w_1^e$ implies that $\frac{\partial s^*}{\partial \mathbf{p}_i^*}=(C^*)^{-1}w_1^*$, where w_i^* are defined in an obvious way. As a consequence of the nonsingularity of C^e, C^* is nonsingular, too. In view of (A.1), C is nonsingular. $\square$

Acknowledgement. The authors thank the referees for their invaluable suggestions and opinions.

References

GILL, R. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* 11, 49-58.

GROENEBOOM, P. AND WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Birkhäuser Verlag, Basel.

GU, M.G. AND ZHANG, C-H. (1993). Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.* 21, 611-624.

KAPLAN, E. L. AND MEIER, P. (1958). Nonparametric estimation from incomplete observations. *JASA.* 53, 457-481.

KIEFER, J. AND WOLFOWITZ, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Statist.* 27, 887-906.

ODELL, P.M., ANDERSON, K.M. AND D'AGOSTINO, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics,* 48, 951-959.

PETO, R. (1973). Experimental survival curves for interval-censored data. *Applied Statist.* 22, 86-91.

RAO, C. R. (1973). *Linear Sstatistical Inference and its Applications.* Wiley, NY.

SAMUELSEN, S. O. AND KONGERUD, J. (1994). Interval censoring in longitudinal data of respiratory symptoms in aluminium potroom workers: A comparison of methods. *Statist. in Medicine.* 13, 1771-1780.

STUTE, W. AND WANG, J.L. (1993) The strong law under random censorship. *Ann. Statist.* 21, 1591-1607.

TURNBULL, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Roy. Statist. Soc. Ser.* B. 38, 290-295.

YU, Q. AND LI, L. (1994). On the strong consistency of the product-limit estimator. *Sankhya.* A. 56, 416-430.

YU, Q., LI, L. AND WONG, G. (1996). On consistency of the self-consistent estimator of survival functions with interval censored data. (Submitted to *Scan. J. Statist.*, under revision).

YU, Q., SCHICK, A., LI, L. AND WONG, G.W.C. (1998). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Prob. and Statist. Lett.* 223-228.

QIQING YU
DEPARTMENT OF MATHEMATICAL SCIENCES
STATE UNIVERSITY OF NEW YORK
BINGHAMTON, NY 13902
USA
e-mail: qyu@math.binghamton.edu

GEORGE Y.C. WONG
STRANG CANCER PREVENTION CENTER
NY 10021
USA

LINXIONG LI
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF NEW ORLEANS
LA 70148
USA
e-mail: lli@math.uno.edu

STATISTICS & PROBABILITY LETTERS

Asymptotic properties of the GMLE with case 2 interval-censored data

Qiqing Yu[a,*,1], Anton Schick[a], Linxiong Li[b,2], George Y.C. Wong[c,3]

[a] Department of Mathematical Sciences, Binghamton University, Box 6000, Binghamton, NY 13902, USA
[b] Department of Mathematics, University of New Orleans, LA 70148, USA
[c] Strang Cancer Prevention Center, Cornell University Medical School, NY 10021, USA

Asymptotic properties of the GMLE with case 2 interval-censored data

Qiqing Yu[a,*,1], Anton Schick[a], Linxiong Li[b,2], George Y.C. Wong[c,3]

[a] *Department of Mathematical Sciences, Binghamton University, Box 6000, Binghamton, NY 13902, USA*
[b] *Department of Mathematics, University of New Orleans, LA 70148, USA*
[c] *Strang Cancer Prevention Center, Cornell University Medical School, NY 10021, USA*

Abstract

In case 2 interval censoring the random survival time X of interest is not directly observable, but only known to have occurred before Y, between Y and Z, or after Z, where (Y, Z) is a pair of observable inspection times such that $Y < Z$. We consider the large sample properties of the generalized maximum likelihood estimator (GMLE) of the distribution function of X with case 2 interval-censored data in which the inspection times are discrete random variables. We prove the strong consistency of the GMLE at the support points of the inspection times and establish its asymptotic normality in the case of only finite many support points. © 1998 Elsevier Science B.V. All rights reserved

1. Introduction

In many biomedical studies, the random survival time X of interest is never observed and is only known to lie before an inspection time Y, between two consecutive inspection times Y and Z, or after the inspection time Z. This censoring scheme is referred to as case 2 interval censoring. Examples can be found in cancer studies (Finkelstein and Wolfe, 1985) and AIDS studies (Becker and Melbye, 1991; Aragon and Eberly, 1992). We assume throughout that X and (Y, Z) are independent and that $Y < Z$ with probability one. We denote the distribution function of X by F_0 and the joint distribution function of (Y, Z) by G. The available

---

data for the case 2 interval-censorship model are thus

$$(Y_j, Z_j, I[X_j \leqslant Y_j], I[Y_j < X_j \leqslant Z_j]), \quad j = 1, \ldots, n,$$

where $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$ are independent copies of (X, Y, Z) and $I[A]$ is the indicator of the set A.

Groeneboom and Wellner (1992) considered the case 2 interval-censorship model with continuous F_0 and absolutely continuous G. They proposed an iterative convex minorant algorithm to calculate the GMLE and proved the uniform strong consistency of the GMLE. They showed that the estimator of F_0 obtained at the first step of the iterative convex minorant algorithm converges to F_0 at the $(n \log n)^{1/3}$ rate and that its asymptotic distribution is not normal. The asymptotic distribution of the GMLE remains unresolved. There are other approaches to derive the GMLE. They include Peto's (1973) Newton–Raphson algorithm and Turnbull's (1976) self-consistent algorithm.

In this paper, we assume that F_0 is arbitrary, but (Y, Z) is discrete. This assumption is used by several authors (Becker and Melbye, 1991; Finkelstein, 1986). Let

$$\mathscr{A} = \{a \in \mathbb{R}: P(Y = a) + P(Z = a) > 0\}$$

be the set of all possible values of Y or Z. We establish the strong consistency of the GMLE at each point in $\mathscr{A}$. From this we can then infer the uniform strong consistency of the GMLE if F_0 is continuous and $\mathscr{A}$ is dense in $[0, \infty)$. This is done in Section 2.

In Section 3 we consider the case of finite $\mathscr{A}$. We obtain the joint asymptotic normality of the GMLE at the usual $\sqrt{n}$ rate for the points in $\mathscr{A}$ and present a consistent estimator of its asymptotic variance.

2. The consistency of the GMLE

Let $\mathscr{B}$ denote the set of all pairs (a, b) such that $g(a, b) = P(Y = a, Z = b) > 0$. In other words, $\mathscr{B}$ is the set of all possible values of (Y, Z). For $(a, b) \in \mathscr{B}$, let

$$N_n^-(a, b) = \frac{1}{n} \sum_{j=1}^{n} I[X_j \leqslant a, Y_j = a, Z_j = b],$$

$$N_n^0(a, b) = \frac{1}{n} \sum_{j=1}^{n} I[a < X_j \leqslant b, Y_j = a, Z_j = b],$$

$$N_n^+(a, b) = \frac{1}{n} \sum_{j=1}^{n} I[X_j > b, Y_j = a, Z_j = b],$$

$$N_n(a, b) = \frac{1}{n} \sum_{j=1}^{n} I[Y_j = a, Z_j = b].$$

Then the generalized likelihood is given by

$$\Lambda_n(F) = \prod_{(a,b) \in \mathscr{B}} F(a)^{nN_n^-(a,b)} [F(b) - F(a)]^{nN_n^0(a,b)} [1 - F(b)]^{nN_n^+(a,b)}$$

and the normalized generalized log-likelihood is

$$\mathscr{L}_n(F) = \sum_{(a,b) \in \mathscr{B}} \{N_n^-(ab) \log[F(a)] + N_n^0(a, b) \log[F(b) - F(a)] + N_n^+(a, b) \log[1 - F(b)]\}.$$

Here and below we interpret $0 \log 0 = 0$ and $\log 0 = -\infty$. In the above we let F range over the set $\mathscr{F}^*$ of all subdistribution functions. A function F_1 is called a subdistribution function if $F_1 = aF$ for some distribution function F and a number $a \in [0, 1]$. Note that $\Lambda_n(F)$ and $\mathscr{L}_n(F)$ depend on F only through the values of F at the points $a \in \mathscr{A}$. Thus there exists no unique maximizer of $\Lambda_n(F)$ over the set $\mathscr{F}^*$. However, there exists a unique maximizer $\hat{F}_n$ of $\Lambda_n(F)$ over the set $\mathscr{F}^*$ which satisfies $\hat{F}_n(x) = \sup\{\hat{F}_n(a): a \leqslant x, \sum_{j=1}^{n}(I[Y_j = a] + I[Z_j = a]) > 0\}$ for all $x \in \mathbb{R}$. Here we interpret the supremum of the empty set as 0. We call $\hat{F}_n$ the GMLE of F_0.

Theorem 2.1. *The GMLE $\hat{F}_n$ satisfies $\hat{F}_n(a) \to F_0(a)$ almost surely for all $a \in \mathscr{A}$.*

Proof. Verify that

$$L(F) := E(\mathscr{L}_n(F)) = \sum_{(a,b) \in \mathscr{B}} g(a, b) h_{a,b}(F)$$

with

$$h_{a,b}(F) = F_0(a) \log[F(a)] + [F_0(b) - F_0(a)] \log[F(b) - F(a)] + [1 - F_0(b)] \log[1 - F(b)].$$

It is easy to check that the expression $h_{a,b}(F)$ is maximized by a nondecreasing function into $[0,1]$ F if and only if $F(a) = F_0(a)$ and $F(b) = F_0(b)$. Thus, F_0 maximizes $L(F)$ and any other nondecreasing function into $[0,1]$ that maximizes $L(F)$ coincides with F_0 at the points in $\mathscr{A}$.

Note that $\mathscr{L}_n(F_0) = \frac{1}{n} \sum_{j=1}^{n} \psi(X_j, Y_j, Z_j)$, where ψ is the map defined by

$$\psi(x, y, z) = I[x \leqslant y] \log(F_0(y)) + I[y < x \leqslant z] \log(F_0(z) - F_0(y)) + I[z < x] \log(1 - F_0(z)).$$

Thus, it follows from the SLLN that $\mathscr{L}_n(F_0) \to L(F_0)$ almost surely. By the definition of the GMLE, $\mathscr{L}_n(\hat{F}_n) \geqslant \mathscr{L}_n(F_0)$. Consequently,

$$\liminf_{n \to \infty} \mathscr{L}_n(\hat{F}_n) \geqslant \liminf_{n \to \infty} \mathscr{L}_n(F_0) = L(F_0) \quad \text{almost surely.}$$

Let Ω' denote the event on which $\liminf_{n \to \infty} \mathscr{L}_n(\hat{F}_n) \geqslant L(F_0)$ and, for each $(a, b) \in \mathscr{B}$, $N_n^-(a, b) \to F_0(a) g(a, b)$, $\sup_n N_n^-(a, b) = 0$ if $F_0(a) = 0$, $N_n^0(a, b) \to (F_0(b) - F_0(a)) g(a, b)$, $\sup_n N_n^0(a, b) = 0$ if $F_0(b) = F_0(a)$, $N^+(a, b) \to (1 - F_0(b)) g(a, b)$ and $\sup_n N_n^+(a, b) = 0$ if $F_0(b) = 1$. Fix an $\omega \in \Omega'$. Let the function F^* be a limit point of $\hat{F}_n(\cdot, \omega)$ in the sense that $\hat{F}_{k_n}(a, \omega) \to F^*(a)$ for all $a \in \mathscr{A}$ and for some sequence $\{k_n\}$ of positive integers tending to infinity. We now show that

$$L(F^*) \geqslant L(F_0).$$

Let $x_{k_n}(a, b)$ denote the value of the random variable

$$N_{k_n}^-(a, b) \log(\hat{F}_{k_n}(a)) + N_{k_n}^0(a, b) \log(\hat{F}_{k_n}(b) - \hat{F}_{k_n}(a)) + N_{k_n}^+(a, b) \log(1 - \hat{F}_{k_n}(b))$$

at the point ω. Thus, by the definition of Ω',

$$\liminf_{n \to \infty} \sum_{(a,b) \in \mathscr{B}} x_{k_n}(a, b) \geqslant L(F_0)$$

and

$$x_{k_n}(a, b) \to g(a, b) h_{a,b}(F^*)$$

for each $(a,b) \in \mathscr{B}$. Note also that $x_{k_n}(a,b) \leqslant 0$ for all $(a,b) \in \mathscr{B}$. Thus an application of Fatou's Lemma yields

$$\limsup_{n \to \infty} \sum_{(a,b) \in \mathscr{B}} x_{k_n}(a,b) = -\liminf_{n \to \infty} \sum_{(a,b) \in \mathscr{B}} -x_{k_n}(a,b)$$

$$\leqslant -\sum_{(a,b) \in \mathscr{B}} \liminf_{n \to \infty}(-x_{k_n}(a,b))$$

$$= \sum_{(a,b) \in \mathscr{B}} g(a,b) h_{a,b}(F^*)$$

$$= L(F^*).$$

Combining the above yields $L(F_0) \leqslant L(F^*)$. As F_0 maximizes L, we can conclude that $L(F^*) = L(F_0)$ and therefore $F^*(a) = F_0(a)$ for all $a \in \mathscr{A}$. Since ω was arbitrary and Ω' has probability one, we can infer the desired result. $\square$

If $\mathscr{A}$ is a finite set, then it follows from the theorem that the GMLE is uniformly strongly consistent on $\mathscr{A}$. For arbitrary $\mathscr{A}$, the uniform strong consistency of the GMLE requires additional assumptions. The proofs of the following corollary and theorem are similar to Yu et al. (1998) and are thus omitted here.

Corollary 2.2. *Suppose that $\mathscr{A}$ is a closed set. Assume that $F_0(a-) = F_0(a)$ for every $a \in \mathscr{A}$ for which there is a sequence of points $\{a_i\}_{i \geqslant 1} \subset \mathscr{A}$ such that $a_i \uparrow a$. Then the GMLE is uniformly strongly consistent on $\mathscr{A}$, i.e., $\sup_{a \in \mathscr{A}} |\hat{F}_n(a) - F_0(a)| \to 0$ almost surely.*

We call a number x a *point of increase* of F_0 if either $F_0(x) < F_0(y)$ for all $y > x$ or $F_0(y) < F_0(x)$ for all $y < x$.

Theorem 2.3. *Suppose that F_0 is continuous and the closure of $\mathscr{A}$ contains the set of all points of increase of F_0. Then the GMLE is uniformly strongly consistent, i.e., $\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)| \to 0$ almost surely.*

3. The asymptotic normality of the GMLE

In this section we shall obtain the asymptotic normality of the GMLE under the assumption that $\mathscr{A}$ contains finitely many elements and

$$0 < F_0(a) < F_0(b) < 1 \quad \text{for all } a,b \text{ in } \mathscr{A} \text{ such that } a < b.$$

Note that under the current assumption the standard method for finite parametric models can be used.

Let $\mathscr{F}$ denote the set of all distribution functions F which satisfy $0 < F(a) < F(b) < 1$ for all a,b in $\mathscr{A}$ with $a < b$. For $F \in \mathscr{F}$ and $a \in \mathscr{A}$, let

$$\mathscr{L}_{n,a}(F) = \sum_{b: (a,b) \in \mathscr{B}} \left(\frac{N_n^-(a,b)}{F(a)} - \frac{N_n^0(a,b)}{F(b) - F(a)} \right) + \sum_{c: (c,a) \in \mathscr{B}} \left(\frac{N_n^0(a,b)}{F(a) - F(c)} - \frac{N_n^+(c,a)}{1 - F(a)} \right),$$

$$\mathscr{L}_{n,a,a}(F) = -\sum_{b: (a,b) \in \mathscr{B}} \left(\frac{N_n^-(a,b)}{F^2(a)} + \frac{N_n^0(a,b)}{(F(b) - F(a))^2} \right)$$

$$- \sum_{c: (c,a) \in \mathscr{B}} \left(\frac{N_n^0(a,b)}{(F(a) - F(c))^2} + \frac{N_n^+(c,a)}{(1 - F(a))^2} \right)$$

and

$$\mathscr{L}_{n,a,b}(F) = \mathscr{L}_{n,b,a}(F) = \frac{N_n^0(a,b)}{(F(b) - F(a))^2}, \quad a, b \in \mathscr{A}, a < b.$$

Then

$$\mathscr{L}_{n,a}(F) = \frac{\partial \mathscr{L}_n(F)}{\partial F(a)} \quad \text{and} \quad \mathscr{L}_{n,a,b}(F) = \mathscr{L}_{n,b,a}(F) = \frac{\partial^2 \mathscr{L}_n(F)}{\partial F(a)\partial F(b)}, \quad a, b \in \mathscr{A}.$$

Let $a_1 < a_2 < \cdots < a_m$ denote the elements of $\mathscr{A}$. For $F \in \mathscr{F}$, let $\dot{\mathscr{L}}_n(F)$ denote the m-dimensional column vector with entries $(\dot{\mathscr{L}}_n(F))_i = \mathscr{L}_{n,a_i}(F)$, $i = 1, \ldots, m$, and $\ddot{\mathscr{L}}_n(F)$ denote the $m \times m$ matrix with entries

$$(\ddot{\mathscr{L}}_n(F))_{ij} = \mathscr{L}_{n,a_i,a_j}(F), \quad i, j = 1, \ldots, m.$$

Finally, set

$$J = E[\dot{\mathscr{L}}_n(F_0)(\dot{\mathscr{L}}_n(F_0))^{\mathrm{T}}] = -E[\ddot{\mathscr{L}}_n(F_0)].$$

The matrix J is positive definite since

$$J = D + \sum_{1 \leq i < j \leq m} \frac{g(a_i, a_j)}{F_0(a_j) - F_0(a_i)}(e_i - e_j)(e_i - e_j)^{\mathrm{T}},$$

where D is the diagonal matrix with positive diagonal elements

$$d_{ii} = \frac{P\{Y = a_i\}}{F_0(a_i)} + \frac{P\{Z = a_i\}}{1 - F_0(a_i)}, \quad i = 1, \ldots, m,$$

and $e_1, \ldots, e_m$ denote the standard basis in $\mathbb{R}^m$. It is easy to verify that

$$\ddot{\mathscr{L}}_n(\hat{F}_n) \to E[\ddot{\mathscr{L}}_n(F_0)] = -J.$$

It thus follows that on the event $\{\hat{F}_n \in \mathscr{F}\}$

$$0 = \dot{\mathscr{L}}_n(\hat{F}_n) = \dot{\mathscr{L}}_n(F_0) - J\Delta_n + \mathrm{o}_{\mathrm{p}}(\|\Delta_n\|),$$

where Δ_n is the m-dimensional column vector with entries $\hat{F}_n(a_i) - F_0(a_i)$, $i = 1, \ldots, m$. It follows from the CLT that $n^{1/2} \dot{\mathscr{L}}_n(F_0)$ is asymptotically normal with mean 0 and dispersion matrix J. This shows that $\Delta_n = J^{-1} \dot{\mathscr{L}}_n(F_0) + \mathrm{o}_{\mathrm{p}}(n^{-1/2})$. Thus, we have the following result.

**Theorem 3.1.** *Suppose F_0 belongs to $\mathscr{F}$. Then*

$$n^{1/2} \begin{pmatrix} \hat{F}_n(a_1) - F_0(a_1) \\ \vdots \\ \hat{F}_n(a_m) - F_0(a_m) \end{pmatrix}$$

is asymptotically normal with mean 0 and dispersion matrix J^{-1}. A strongly consistent estimator of J is given by $-\ddot{\mathscr{L}}_n(\hat{F}_n)$.

References

Aragon, J., Eberly, D., 1992. On convergence of convex minorant algorithms for distribution estimation with interval-censored data. J. Comput. Graphical Statist. 1, 129–140.

Becker, N.G., Melbye, M., 1991. Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity. Austral. J. Statist. 33, 125–133.

Finkelstein, D.M., 1986. A proportional hazards model for interval-censored failure time data. Biometrics 42, 845–854.

Finkelstein, D.M., Wolfe, R.A., 1985. A semiparametric model for regression analysis of interval-censored failure time data. Biometrics 41, 933–945.

Groeneboom, P., Wellner, J.A., 1992. Information Bounds and Nonparametric Maximum Likelihood Estimation. Birkhäuser, Basel.

Peto, R., 1973. Experimental survival curve for interval-censored data. Appl. Statist. 22, 86–91.

Turnbull, B.W., 1976. The empirical distribution function with arbitrary grouped, censored and truncated data. J. Roy. Statist. Soc. Ser. B 38, 290–295.

Yu, Q., Schick, A., Li, L., Wong, G., 1998. Estimation of a survival function with case 1 interval-censored data, to appear in Canad. J. Statist.
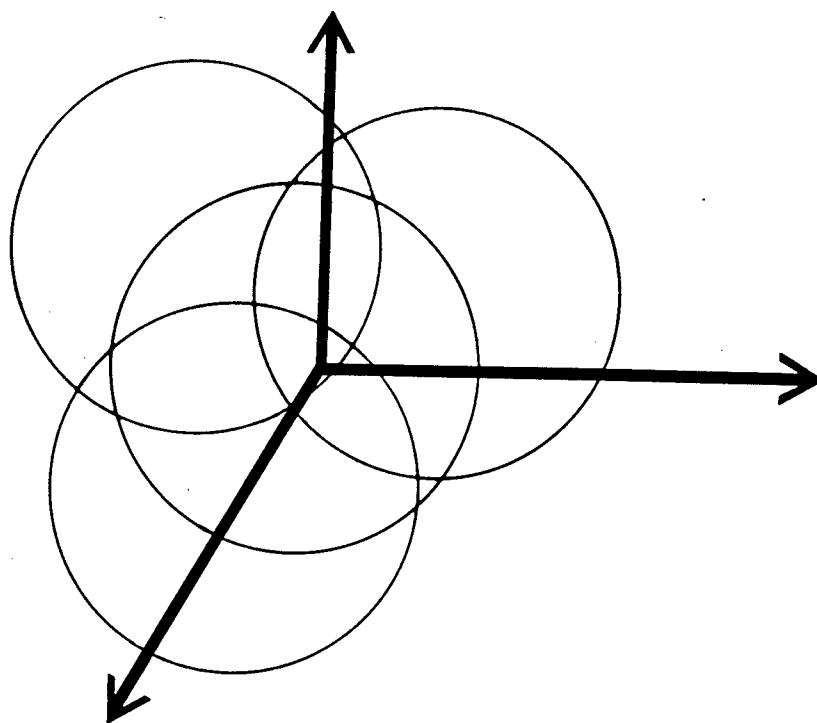
communications in statistics
THEORY AND METHODS

COMMUNICATIONS IN STATISTICS

Part A: Theory and Methods

Communications in Statistics is a multi-part journal.

Part A: Theory and Methods focuses primarily on new applications of known statistical methods to actual problems in industry and government, and has a strong mathematical orientation to statistical studies. In addition, **Part A** also offers communications that discuss practical problems with only ad hoc solutions or none at all; in either case, where there is a difference of opinion on particular techniques, all parties involved vigorously debate the issue with thought-provoking commentary.

For subscription information write the Promotion Department:

MARCEL DEKKER, INC.
270 MADISON AVENUE
NEW YORK, NEW YORK 10016

COMMUN. STATIST.—THEORY METH., 27(6), 1547–1556 (1998)

CONSISTENCY OF SELF-CONSISTENT ESTIMATORS
OF A DISCRETE DISTRIBUTION FUNCTION
WITH BIVARIATE RIGHT-CENSORED DATA

Qiqing Yu and George Y. C. Wong

Math. Dept., SUNY at Binghamton, NY 13902, USA,

and

Strang Cancer Prevention Center, 428 E 72nd Street, NY 10021, USA

Key words : nonparametric MLE, bivariate survival analysis, uniform strong consistency.

ABSTRACT.

The asymptotic properties of the nonparametric maximum likelihood estimator and other estimators of a joint distribution function F of a bivariate random vector X with right-censored data have been studied by several authors. Among others, an important assumption made in their studies is that X lives on a rectangle region $[0, a] \times [0, b]$ which can be observed. However, in many follow-up studies, $a = b = L$ is the length of the study period and X lives on a region larger than $[0, L] \times [0, L]$. Thus it is of interest to study whether the asymptotic results established by these authors are still valid without that restriction. In this direction, we established the strong consistency of self-consistent estimators of a discrete distribution function.

1. INTRODUCTION

We consider the uniform strong consistency of self-consistent estimators (SCE) of a joint distribution function of a discrete bivariate random vector with right-censored data.

Specifically, we consider the following bivariate right censorship model: Let $X = (X_1, X_2)$ be a random survival vector with a joint distribution function $F(\mathbf{x})$ $(= P\{X \le \mathbf{x}\})$, where $\mathbf{x} = (x_1, x_2)$. Here and after, by $X \le \mathbf{x}$, $X \ge \mathbf{x}$

1547

Copyright © 1998 by Marcel Dekker, Inc.

and $X > x$, we mean the inequalities hold componentwise, e.g., $X_i \leq x_i$ for all i. Let $Y = (Y_1, Y_2)$ be a random censoring vector with a joint distribution function $G(y)$, where $y = (y_1, y_2)$. Assume that X and Y are independent. We consider the estimation of $F(x)$ (or the survival function $S(x) = P\{X > x\}$). Let (X_{i1}, X_{i2}), $i = 1, ..., n$, be i.i.d. copies of X and (Y_{i1}, Y_{i2}), $i = 1, ..., n$, i.i.d. copies of Y. We observe $((\min\{X_{ij}, Y_{ij}\}, 1(X_{ij} \leq Y_{ij})), j = 1, 2)$, $i = 1, 2, ..., n$, where $1(A)$ is the indicator function of a set A. Let $\tau_i = \inf\{x : P(X_i \leq x) = 1$ or $P(Y_i \leq x) = 1\}$, $i = 1, 2$, and $\tau = (\tau_1, \tau_2)$.

Multivariate survival data arise naturally in eye diseases research, twin studies and studies of family history. For examples, we refer to Campbell (1981), Hanley and Parnes (1983), and Clayton and Cuzick (1985).

Prentice and Cai (1992) consider the estimation of the covariance and the multivariate survival function using censored multivariate failure time data. Lin and Ying (1993) propose a simple nonparametric estimate of the bivariate survival function under univariate censoring and studied its asymptotic properties. Campbell (1981), Tsai et al (1986), Pruitt (1991) and Gill (1992) all study estimators of a multivariate survival function.

Dabrowska (1988) introduces the Kaplan-Meier estimate on the plane. She proves its consistency and Gill (1992) discusses its asymptotic normality. Note that in general the Kaplan-Meier estimate on the plane is not a nonparametric maximum likelihood estimator (NPMLE) of a survival function with bivariate right-censored data.

Hanley and Parnes (1983) study the NPMLE. The asymptotic normality of the NP MLE $\hat{F}$ of F has been addressed under the homogeneous right-censoring case (see Hanley and Parnes (1983) and Lin and Ying (1993)), since a closed form solution for the NPMLE of $\hat{F}$ is available. Note that the NPMLE of F is not unique for the continuous bivariate random vector. Thus it is convenient to have a closed form solution for the NPMLE. Homogeneous right-censored data may arise under univariate censoring, i.e, $Y_1 = Y_2$ w.p.1, (see Lin and Yin (1993)). In general, bivariate right-censored data are heterogeneous (Hanley and Parnes (1983)).

It is well known that if F is continuous then the NPMLE is not consistent in general (see Tsai et al. (1986)). Thus it is only relevant to study asymptotic properties of the NPMLE under the assumption that F is discrete or to investigate a certain consistent or efficient estimator of F.

Van der Laan (1996) proposes a repaired NPMLE. Under the assumption that

$$P\{X \leq \tau\} = 1, \ P\{X = \tau\} > 0 \text{ and } P\{Y > \tau\} > 0; \tag{1.1}$$

and additional assumptions, he shows that his estimator is asymptotically efficient. Other estimators of F based on bivariate right-censored data are discussed in his paper.

Note that the NPMLE is also an SCE (Tsai and Crowley (1985)) but SCEs may not be unique (see Example 1 in Section 2). Furthermore, van der Laan's estimator (the repaired NPMLE) is not a NPMLE in general even under the assumption that

$$X \text{ is a discrete random vector} \tag{1.2}$$

(see also Example 1). Thus his result does not imply that the NPMLE is consistent under (1.1) and (1.2). The asymptotic properties of the NPMLE and SCEs with or without assumption (1.1) remain an open question.

In many situations, assumption (1.1) is not satisfied. In fact, in many medical follow-up studies, studies have to stop at a certain time t_o and the survival times beyond t_o will always be censored. If $F(t_o, t_o) > 0$, then (1.1) fails to hold. Thus it is of interests to study the asymptotic properties of SCEs when (1.1) is not valid. In this paper, we assume

$$P\{Y = \tau\} > 0 \text{ if } P(X \leq \tau) < 1. \tag{1.3}$$

Note that van der Laan's efficient results on the repaired NPMLE based on his slightly reduced bivariate data depend on the fact that the NPMLE is consistent if F is discrete and assumption (1.1) holds. Once we can show that the SCE is consistent under alternative assumptions in the case that F is discrete, we may be able to use his technique to show that a modified SCE analog to his repaired NPMLE is also efficient under the new assumptions. However, in view of the lengthy derivation in his paper, we will not investigate this problem here.

In section 2, we give notations used in the paper. In section 3, we prove the uniform strong consistency on the domain $\{x \leq \tau\}$ under assumptions (1.2) and (1.3).

2. NOTATIONS

Hereafter, denote $[0, \tau] = [0, \tau_1] \times [0, \tau_2]$ and denote $[0, x]$ and $[\tau, +\infty)$ in a similar manner. The observable random vector $(\min\{X_i, Y_i\}, 1(X_i \leq Y_i), i = 1, 2)$

is equivalent to an observable random set which takes on values of the following forms: assuming $X = \bar{x}$ and $Y = \bar{y}$,

$$I = \begin{cases} \{(a,b): a = x_1,\ b = x_2\} & \text{if } X \text{ is uncensored,} \\ \{(a,b): a = x_1,\ b > y_2\} & \text{if only } X_2 \text{ is censored,} \\ \{(a,b): a > y_1,\ b = x_2\} & \text{if only } X_1 \text{ is censored,} \\ \{(a,b): a > y_1,\ b > y_2\} & \text{if both } X_1 \text{ and } X_2 \text{ are censored.} \end{cases} \quad (2.1)$$

As in Hanley and Parnes (1983), we denote these four forms of sets by

$$(x_1, x_2),\ (x_1, y_2+),\ (y_1+, x_2) \text{ and } (y_1+, y_2+), \quad (2.2)$$

respectively. We can define an observable random set I which takes on values Is as in (2.1). Let $V = (V_1, V_2, V_3, V_4)$ be the lower-left and upper-right vertexes of the random rectangle I and Q the c.d.f. of V. A solution H_n to the (integral) equation

$$H_n(x) = \int \frac{\mu_{H_n}(I \cap [0, x])}{\mu_{H_n}(I)} d\hat{Q}(v_1), \quad H_n \in \Theta, \quad (2.3)$$

is called a self-consistent estimator of F, where μ_{H_n} is the measure induced by H_n, v_1 is a realization of V, $\hat{Q}$ is the empirical version of Q and Θ is the collection of all functions H from $[-\infty, +\infty] \times [-\infty, +\infty]$ to $[0, 1]$ such that H can induce a measure satisfying

$$\mu_H((x, a] \times (y, b]) = H(a, b) + H(x, y) - H(x, b) - H(a, y) \geq 0 \ \forall\ (x, y) \leq (a, b), \quad (2.4)$$

$H(+\infty, +\infty) = 1$ and $H(-\infty, -\infty) = 0$.

Given a sample, define a maximal intersection (MI) A, of observed I_i's as a nonempty finite intersection of the I_is such that for each i we have $A \cap I_i = \emptyset$ or A. Let $A_1, A_2, \ldots$, be the possible distinct MI's. These A_j are subsets of the forms in (2.2). Verify that the NPMLE $\hat{F}(x)$ does not assign mass to a set which is disjoint with each MI and is not uniquely determined for x in an MI which is a nonsingleton set. In order to uniquely determine SCEs $H_n(x)$ in nonsingleton MI's, hereafter, we refer SCEs H_n to those which put mass only to the upper-right vertexes of MI's. Let A_n be the collection of MI's based on the sample of size n and A be the set of point x such that $\mu_F(\{x\}) > 0$.

The following is an example that the solution to (2.3) may not be unique.

Example 1. Suppose that there are 6 observations: (3, 5), (5, 3), (5, 2+), (2+, 5), (2+, 5), (2+, 5). Then there are 2 SCEs, say $\hat{F}_1$ and $\hat{F}_2$, where $\hat{F}_1$ puts mass 1/2, 1/4 and 1/4 to the points (3, 5), (5, 5) and (5, 3), respectively, and $\hat{F}_2$ puts mass

2/3 and 1/3 to (3, 5) and (5, 3), respectively. Verify that $\hat{F}_1$ is the NPMLE and $\hat{F}_2$ is van der Laan's estimator as his estimator does not put mass on censored points unless the points belong to some observed rectangle I_i which does not contain any uncensored observations.

3. ALMOST SURE LIMIT OF SCES

We shall establish the strong consistency of the SCE H_n under assumptions (1.2) and (1.3). An $H \in \Theta$ is called a *pointwise limit* of H_n if there exists a subsequence $\{H_{n_k}\}$ of $\{H_n\}$ such that $H_{n_k}(x)$ converges at each point on $[-\infty, +\infty]^2$. The existence of such limits is ensured by the following lemma.

Lemma 3.1 *Under assumptions (1.2) and (1.3), for each ω in the sample space, each subsequence of SCEs $\{H_n\}$ has a pointwise limit.*

Proof. It suffices to show that each $\{H_{n_k}(\omega)(x)\}$ has a convergent subsequence. For simplicity, we only show that $\{H_n(\omega)\}$ has a pointwise limit and we suppress the ω in $H_n(\omega)$.

Let B_n be the set of all upper-right vertexes of A's in A_n, then $\cup_n B_n$ is countable. Let B be a countable dense subset of $[-\infty, +\infty]^2$ such that it contains $\cup_n B_n$ and all points whose coordinates are rational. By Helly's selection theorem, there exist a function H and a subsequence H_{n_k} of H_n such that $H_{n_k}(x)$ converges to $H(x)$ for all $x \in B$. Without loss of generality (WLOG), we can assume $\{H_n\}$ is such a subsequence. Note that $F_1(x) = \varprojlim_{n\to\infty} H_n(x)$ and $F_2(x) = \varlimsup_{n\to\infty} H_n(x)$ exist for each $x \in [-\infty, +\infty]^2$, and $F_1(x) = F_2(x)$ for all $x \in B$. We shall show that if $x \notin B$, then $F_1(x) = F_2(x)$ and thus $H_n(x)$ converges too.

For a given $x \notin B$, we shall show that x is a continuity point of F_i, $i = 1, 2$. For $z = (z_1, z_2)$ and $y = (y_1, y_2)$, denote $[y, z] = [y_1, z_1] \times [y_2, z_2]$. For each n, there exist $y_n, z_n \in B$ such that $y_n < z_n$, x belongs to the interior of the set $[y_n, z_n]$ and $[y_n, z_n] \cap B_n = \emptyset$. Thus $\mu_{H_n}([y_n, z_n]) = 0$ for all n as μ_{H_n} only puts mass on the set B_n by assumption. Then $\mu_{F_i}(\{x\}) \leq \varlimsup_{n\to\infty} \mu_{H_n}([y_n, z_n]) = 0$, $i = 1, 2$. Consequently, x is a continuity point of F_i and thus is uniquely determined by $\{F_i(y):\ y \in B\}$ as B is dense. Then $F_1(x) = F_2(x)$ as $F_1(y) = F_2(y)$ for all $y \in B$. Since x is arbitrary, H_n converges on $[-\infty, +\infty]^2$. This concludes the proof of the lemma. $\square$

Let H be a pointwise limit of H_n and D the collection of all singleton sets. Since $\frac{\mu_{H_n}(I \cap [0, x])}{\mu_{H_n}(I)}$ is bounded by 1 a.e. $d\hat{Q}_n$, it follows from the SLLN and bounded convergence theorem (BCT) that H is a solution to

$$H(x) = \left[\int_{I \notin D} \frac{\mu_H(I \cap [0,x])}{\mu_H(I)} + \int_{I \subset [0,x], I \in D} 1\right] dQ(v_I), \quad H \in \Theta. \tag{3.1}$$

Solutions to (3.1) may not be unique in general. For example, let X take values (2,3) and (3,2) with probabilities 1/2 and 1/2, respectively, and let Y take values (4,1) and (1,4) with probabilities 1/2 and 1/2, respectively. Then there are at least two solutions to (3.1): H_1 puts mass 1/2 and 1/2 at points (3,2) and (2,3), respectively; H_2 puts mass 1/2 and 1/2 at points (2,2) and (3,3), respectively. However, under assumptions (1.2) and (1.3), the solution is unique as stated in the following theorem.

Theorem 3.1. *Under assumptions (1.2) and (1.3), if H is a solution to (3.1), then (1) $H(x) = F(x)$ for all $x \leq \tau$; (2) $\mu_H((x, \tau_2+)) = \mu_F((x, \tau_2+))$, $x \leq \tau_1$; (3) $\mu_H((\tau_1+, x)) = \mu_F((\tau_1+, x))$, $x \leq \tau_2$; and (4) $\mu_H((\tau_1+, \tau_2+)) = \mu_F((\tau_1+, \tau_2+))$.*

Let H be a solution to (3.1) and $x = (x_1, x_2) \leq \tau$, Theorem 3.1 implies that $H(x)$, $H(x_1-, x_2)$, $H(x_1, x_2-)$ and $H(x_1-, x_2-)$ are uniquely determined and H is right continuous. Under the discrete assumption (1.2), each subsequence of $\{H_n\}$ has a pointwise limit by Lemma 3.1. Since H_n is an SCE, each pointwise limit satisfies (3.1) by the BCT and the pointwise limit is unique by Theorem 3.1. Consequently, $H_n(x)$, $H_n(x_1-, x_2)$, $H_n(x_1, x_2-)$, and $H_n(x_1-, x_2-)$ converge almost surely to $F(x)$, $F(x_1-, x_2)$, $F(x_1, x_2-)$ and $F(x_1-, x_2-)$, respectively. This yields the pointwise strong consistency as well as the uniform strong consistency on the region $\{x \leq \tau\}$, as $F \in \Theta$ and is right continuous.

Theorem 3.2. *Under the same assumptions as in Theorem 3.1, the SCE H_n satisfies $\lim_{n \to \infty} \sup_{x \leq \tau} |H_n(x) - F(x)| = 0$ with probability one.*

Proof of Theorem 3.1. Let $\mu_i(t) = [F_{X_i}(t) + G_{Y_i}(t)]/2$, $i = 1, 2$, where F_{X_i} and G_{Y_i} are marginal distribution functions. The proof for the case $F_{X_i}(\tau_i) = 1$ for at least one i is similar to the one for the case $F(\tau) < 1$. For simplicity, we assume that $F(\tau) < 1$ and $\tau_1, \tau_2 > 0$ in the proof.

Given a positive integer α, denote $B_{\alpha,i}$ the collection of $+\infty$ and all $100 \times j2^{-\alpha}$, $j = 0, 1, \ldots, 2^\alpha$, percentiles of μ_i which are contained in $[0, \tau_i]$, $i = 1, 2$. Let $B_\alpha = B_{\alpha,1} \times B_{\alpha,2}$. Let $a_1 < \cdots < a_h < \cdots$ be all elements of $B_{\alpha,1}$ and let $b_1 < \cdots < b_j < \cdots$ be all elements of $B_{\alpha,2}$. Define discrete random variables

$$Y_1^\alpha = \sum_{j \geq 1} a_j 1(Y_1 \in [a_j, a_{j+1})) \quad \text{and} \quad Y_2^\alpha = \sum_{j \geq 1} b_j 1(Y_2 \in [b_j, b_{j+1})). \tag{3.2}$$

Define a random rectangle $\mathcal{I}_\alpha = \mathcal{I}_{1,X,Y^\alpha} \times \mathcal{I}_{2,X,Y^\alpha}$, where

$$\mathcal{I}_{1,X,Y^\alpha} = \begin{cases} [a_h, a_h] & \text{if } X_1 = a_h \leq Y_1^\alpha \text{ for some } h \geq 1, \\ (a_h, a_{h+1}) & \text{if } a_h < X_1 < a_{h+1} \leq Y_1^\alpha \text{ for some } h \geq 1, \\ (a_h, +\infty) & \text{if } X_1 > a_h = Y_1^\alpha \text{ for some } h \geq 1, \end{cases} \tag{3.3}$$

and

$$\mathcal{I}_{2,X,Y^\alpha} = \begin{cases} [b_j, b_j] & \text{if } X_2 = b_j \leq Y_2^\alpha \text{ for some } j \geq 1, \\ (b_j, b_{j+1}) & \text{if } b_j < X_2 < b_{j+1} \leq Y_2^\alpha \text{ for some } j \geq 1, \\ (b_j, +\infty) & \text{if } X_2 > b_j = Y_2^\alpha \text{ for some } j \geq 1. \end{cases}$$

Denote the coordinates of the lower-left and upper-right vertexes of the random rectangle $\mathcal{I}_\alpha$ by $V^\alpha = (V_1^\alpha, V_2^\alpha, V_3^\alpha, V_4^\alpha)$. Denote the joint distribution function of the extended random vector V^α by Q_α. It is obvious by construction that μ_{Q_α} converges setwisely to μ_Q as $\alpha \to \infty$. Let $I_{\alpha,h}$s be distinct realizations of $\mathcal{I}_\alpha$. Let $q_{\alpha,h} = P(\mathcal{I}_\alpha = I_{\alpha,h})$. For each $H \in \Theta$, let

$$\psi(H) = E\{\ln[\mu_H(\mathcal{I}_\alpha)/\mu_F(\mathcal{I}_\alpha)]\} \quad \left(= \sum_h q_{\alpha,h} \ln[\mu_H(I_{\alpha,h})/\mu_F(I_{\alpha,h})]\right),$$

where we define $\ln 0 = -\infty$, $0\ln 0 = 0$ and $0\ln\infty = 0$. We call $\psi(H)$ a limiting point of $\psi_\alpha(H)$ if there exists a subsequence $\psi_{n_j}(H)$ converges to $\psi(H)$.

The proofs of the following two lemmas are given in Yu and Wong (1997).

Lemma 3.2. *Assume (1.2) and (1.3) hold. Suppose $H \in \Theta$. Let $\psi(H)$ be a limiting point of $\{\psi_\alpha(H)\}$. Then $\psi(H) \leq 0$. Moreover, $\psi(H) = 0$ if and only if $H(x) = F(x)$, $H(x_1, +\infty) = F(x_1, +\infty)$ and $H((+\infty, x_2)) = F((+\infty, x_2))$ for all $x = (x_1, x_2) \leq \tau$.*

Let $\mathcal{L}_F = A \cap [0, \tau]$. Denote $\gamma_H(x, y) = \frac{P(X \leq Y, X \in [x, y])}{\mu_H([x, y])}$.

Lemma 3.3. *Assume (1.2) and (1.3) hold. Suppose that H is a solution to (3.1). Then for each $x \in \mathcal{L}_F$ and each $y \leq x$,*

$(L.1)$ $\gamma_H(y, x) \leq 1$; $\qquad (L.2)$ $\int \frac{1(2\tau \in I)}{\mu_H(I)} dQ(v_I) = 1$ if $F(\tau) < 1$:

$(L.3)$ $\int \frac{1((2\tau_1, x_2) \in I)}{\mu_H(I)} dQ(v_I) = 1$ if $x_2 \leq \tau_2$ and $F(+\infty, x_2) > F(\tau_1, x_2)$;

$(L.4)$ $\int \frac{1((x_1, 2\tau_2) \in I)}{\mu_H(I)} dQ(v_I) = 1$ if $x_1 \leq \tau_1$ and $F(x_1, +\infty) > F(x_1, x_2)$;

$(L.5)$ $\int_{I \notin D} \frac{1(x \in I)}{\mu_H(I)} dQ(v_I) + \lim_{y \uparrow x} \gamma_H(y, x) - 1 = 0.$

Suppose that H is a solution of (3.1). We first show statement (1). We shall assume that $H(x) \neq F(x)$ for some $x \leq \tau$ and shall show that it leads to a contradiction.

Let $\psi(H)$ be a limiting point of $\psi_\alpha(H)$. WLOG, assume $\psi_\alpha(H) \to \psi(H)$ as $\alpha \to \infty$. Since $H(t_0) \neq F(t_0)$ for some $t_0 \leq \tau$, $\psi(F) = 0 > \psi(H)$ by Lemma 3.2. Therefore, there exists an integer α_1 such that $\psi_\alpha(F) > \psi_\alpha(H) + \delta$, for all $\alpha \geq \alpha_1$, where $\delta = (0 - \psi(H))/2 > 0$. For each $\alpha \geq \alpha_1$, denote M_α the collection of all MI's induced by all possible realizations of $\mathcal{I}_\alpha$, that is, $I \in M_\alpha \Rightarrow v_I = (v_1, v_2, v_3, v_4)$ where $(v_1, v_3) = (a_i, a_i)$ or (a_i, a_{i+1}) or $(\tau_1, +\infty)$ and $(v_2, v_4) = (b_j, b_j)$ or (b_j, b_{j+1}) or $(\tau_2, +\infty)$. Denote $M_{\infty\alpha}$ the subset of M_α such that the rectangle $I \in M_{\infty\alpha}$ implies that one coordinate of v_I is $+\infty$. Let $p_i = \mu_F(I_{\alpha,i})$, where $I_{\alpha,i}$ is a realization of $\mathcal{I}_\alpha$. Then, for $\alpha \geq \alpha_1$, the above inequality yields

$$\delta \leq - \psi_\alpha(H) + \psi_\alpha(F)$$

$$= \lim_{u\downarrow 0} \frac{\frac{1}{1+u}\psi_\alpha(H) + \frac{u}{1+u}\psi_\alpha(F) - \psi_\alpha(H)}{u}$$

$$\leq \lim_{u\downarrow 0} \frac{\psi_\alpha\left(\frac{1}{1+u}H + \frac{u}{1+u}F\right) - \psi_\alpha(H)}{u} \qquad (\text{since } -\psi_\alpha(H) \text{ is convex})$$

$$= \lim_{u\downarrow 0} \frac{\sum_j q_{\alpha,j}\ln \frac{\mu_{H+uF}(I_{\alpha,j})}{1+u}}{\mu_F(I_{\alpha,j})} - \sum_j q_{\alpha,j}\ln\frac{\mu_H(I_{\alpha,j})}{\mu_F(I_{\alpha,j})}}{u}$$

$$= \lim_{u\downarrow 0} \frac{\sum_j q_{\alpha,j}[\ln(1 + \frac{u\mu_F(I_{\alpha,j})}{\mu_H(I_{\alpha,j})}) - \ln(1 + u)]}{u}$$

$$= \sum_j q_{\alpha,j}\left[\frac{\mu_F(I_{\alpha,j})}{\mu_H(I_{\alpha,j})} - 1\right]$$

$$= \int_{|v_I|=+\infty} \frac{\mu_F(I)}{\mu_H(I)} dQ_\alpha(v_I) + \sum_{I_{\alpha,j} \in M_\alpha \setminus M_{\infty\alpha}} p_j \frac{q_{\alpha,j}}{\mu_H(I_{\alpha,j})} - 1, \qquad (3.4)$$

where $|\cdot|$ is the supremum norm of a vector. For each $I_{\alpha,i}$ in M_α with corresponding $v = (v_1, v_2, v_3, v_4)$, let m_i be a point in the set $I_{\alpha,i}$; specifically, let it be the midpoint of the set $I_{\alpha,i}$ if the set is finite, $(2\tau_1, v_2)$ if $v_3 = +\infty$ and $v_2 = v_4 < +\infty$, $(v_1, 2\tau_2)$ if $v_4 = +\infty$ and $v_1 = v_3 < +\infty$, and 2τ if $v_3 = v_4 = +\infty$. Verify that for a fixed α.

$$\sum_{I_{\alpha,j} \in M_\alpha} p_j = 1 \text{ and}$$

$$+\infty > \lim_{\alpha\to\infty} \sum_{I_{\alpha,j} \in M_\alpha} p_j \int_{|v_I|=+\infty} \frac{1(m_j \in I)}{\mu_H(I)} dQ(v_I) \qquad (\text{by (L.2) - (L.4)})$$

$$\geq \int_{|v_I|=+\infty} \frac{\lim_{\alpha\to\infty} \sum_{I_{\alpha,j} \in M_\alpha} p_j 1(m_j \in I)}{\mu_H(I)} dQ(v_I) \qquad (\text{by Fatou's Lemma})$$

$$= \int_{|v_I|=+\infty} \mu_F(I)/\mu_H(I) dQ(v_I). \qquad (3.5)$$

Since $h_1(v_I) = \mu_F(I)/\mu_H(I)$ is a nonnegative measurable function, (3.5) implies that it is integrable. Verify that for each $I_{\alpha,i} \in M_\alpha$, $q_{\alpha,i} \leq P(X \in I_{\alpha,i}; X \leq Y_\alpha)$ by the definition of Y_α and $I_{\alpha,i}$ (see (3.2) and (3.3)). This, together with (L.1) implies that $|q_{\alpha,i}/\mu_H(I_{\alpha,i})| \leq 1$, and thus $\sum_{I_{\alpha,i} \in M_\alpha \setminus M_{\infty\alpha}} p_i q_{\alpha,i}/\mu_H(I_{\alpha,i})$ converges by the BCT. Then

$$0 < \delta \leq \text{expression (3.4)}$$

$$\leq \lim_{\alpha\to\infty}\left[\int_{|v_I|=+\infty} h_1(v_I) dQ_\alpha(v_I) + \sum_{I_{\alpha,i} \in M_\alpha \setminus M_{\infty\alpha}} \frac{p_i q_{\alpha,i}}{\mu_H(I_{\alpha,i})} - 1\right]$$

$$= \int_{|v_I|=\infty} h_1(v_I) dQ(v_I) + \lim_{\alpha\to\infty} \sum_{I_{\alpha,i} \in M_\alpha \setminus M_{\infty\alpha}} \frac{p_i q_{\alpha,i}}{\mu_H(I_{\alpha,i})} - 1$$

$$\leq \int_{|v_I|=\infty} h_1(v_I) dQ(v_I) + \lim_{\alpha\to\infty} \sum_{I_{\alpha,i} \in M_\alpha \setminus M_{\infty\alpha}} p_i \gamma_H(v_{I_{\alpha,i}}) - 1$$

$$= \lim_{\alpha\to\infty} \sum_{I_{\alpha,i} \in M_\alpha} p_i \int_{|v_I|=\infty} \frac{1(I_{\alpha,j} \subset I)}{\mu_H(I)} dQ(v_I) + \lim_{\alpha\to\infty} \sum_{I_{\alpha,i} \in M_\alpha \setminus M_{\infty\alpha}} p_i \gamma_H(v_{I_{\alpha,i}}) - 1$$

$$\qquad (\text{by (3.5)})$$

$$= \lim_{\alpha\to\infty} \sum_{I_{\alpha,i} \in M_\alpha \setminus M_{\infty\alpha}} p_i\left[\int_{|v_I|=\infty} \frac{1(I_{\alpha,j} \subset I)}{\mu_H(I)} dQ(v_I) + \gamma_H(v_{I_{\alpha,i}}) - 1\right]$$

$$\qquad (\text{by (L.2) - (L.4)})$$

$$= \int_{x \in \mathcal{L}_F}\left[\int_{I \notin D} \frac{1(x \in I)}{\mu_H(I)} dQ(v_I) + \lim_{y\uparrow x} \gamma_H(y, x) - 1\right] dF(x) \qquad (\text{by BCT})$$

$$= 0 \qquad (\text{by (L.5)}).$$

Thus we reach the contradiction $0 < \delta \leq 0$. The contradiction shows that $H(x) = F(x)$ for all $x \leq \tau$ if H is a solution to (3.1).

The proof of statements (2) - (4) are similar and are skipped. This concludes the proof of Theorem 3.1. □

ACKNOWLEDGEMENTS

BIBLIOGRAPHY

Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika.* 68, 417-422.

Clayton, D. G. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). *J. Roy. Statist. Soc. Ser.* A. 148, 82-117.

Dabrowska, D. M. (1988). Kaplan-Meier estimate on the plane. *Ann. Statist.* 16, 1475-1489.

Eastern Cooperative Oncology Group (1973). Phase II study of induction combination chemotherapy and maintenance hormono chemotherapy for metastatic breast carinoma. Eastern Cooperative Oncology Group:905 University Avenue, Madison, Wisconsin.

Gill, R. (1992). Multivariate survival analysis. *Theory Prob. Applic.* 37, 18-31.

Hanley, J. A. and Parnes, M. N. (1983). Nonparametric estimation of a multivariate distribution in the presense of censoring. *Biometrics.* 39. 129-139.

Lin, D.Y. and Ying, Zhiliang. (1993). A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika.* 80, 573-581. 39-69.

Prentice, R. L. and Cai, J. (1992). Covariance and survival function estimation using censored multivariate failure time data. *Biometrika,* 79, 3, 495-512.

Pruitt, R. C. (1991). On negative mass assigned by the bivariate Kaplan-Meier estimator. *ann. Statist.* 19 443-453.

Tsai, W. Y., Leurgans, S. and Crowley, J. (1986). Nonparametric estimation of a bivariate survival function in the presence of censoring. *Ann. Statist.* 14. 1351-1365.

Tsai, W. and Crowley, J. (1985). A large sample study of the generalized maximum likelihood estimators from incomplete data via self - consistency. *Annals of Statistics.* 13, No. 4, 1317-1334.

Van der Laan, M. J. (1996) Efficient estimation in the bivariate censoring model and repairing NPMLE. *Ann. Statist.* 24, 596-627.

Yu, Q. and Wong, G. Y. C. (1997). Technical report on "Consistency of self-consistent estimators of a discrete distribution function with bivariate right-censored data". Math Dept. Binghamton University.

Generalized MLE of a Joint Distribution Function with Multivariate Interval-Censored Data

George Y. C. Wong* and Qiqing Yu†

Strang Cancer Prevention Center, 428 E 72nd Street, New York 10021 and Department of Mathematical Sciences, SUNY, Binghamton

We consider the problem of estimation of a joint distribution function of a multivariate random vector with interval-censored data. The generalized maximum likelihood estimator of the distribution function is studied and its consistency and asymptotic normality are established under the case 2 multivariate interval censorship model and discrete assumptions on the censoring random vectors. © 1999 Academic Press

AMS 1991 subject classifications: 62G05, 62G20.

Key words and phrases: multivariate interval-censored data; asymptotic normality; asymptotic variance; consistent estimate; generalized MLE; multivariate survival analysis.

1. INTRODUCTION

We consider the estimation of a joint distribution function F_0 of a multivariate random vector $\mathbf{X} = (X_1, ..., X_d)$ which is subject to interval censoring. In interval censoring, the value of each coordinate variable X_i may not be directly observable; instead, a pair of extended real numbers L_i and R_i such that $L_i \leqslant X_i \leqslant R_i$ are always observed. The observations L_i and R_i satisfy one of the following four conditions: $L_i = R_i$ (exact), $0 = L_i < R_i$ (left censored), $L_i < R_i = \infty$ (right censored), and $0 < L_i < R_i < \infty$ (strictly interval censored). A d-dimensional interval-censored observation corresponding to $\mathbf{X}$ is represented by the $2d$-dimensional vector $(L_1, R_1, ..., L_d, R_d)$.

Multivariate interval-censored data arise in a variety of life testing situations and biomedical studies. We describe a clinical study in the

---

following example that gives rise to bivariate $(d=2)$ interval-censored data.

EXAMPLE 1.1 (The Italian-American Cataract Study Group (1994)). A total of 1399 persons, between 45 of 79 years of age, who had been identified in a clinic-based case control study were enrolled in a follow-up study between 1985 and 1988. The follow-up study was designed to estimate the rate of incidence and progression of cortical, nuclear, and posterior subcapsular cataracts and to evaluate the usefulness of the Lens Opacities Classification System II in a longitudinal study. Beginning in 1989, follow-up lens photographs were taken and graded at a six-month interval. Patients might skip some visits. Data were obtained from Zeiss slit-lamp and Neitz retroillumination lens photographs at each patient's visit. The exact time that the event of interest occurred was only known to lie within the period between two consecutive visits, or was right censored if by the end of the study the event still had not taken place. Consequently, bivariate interval-censored data were encountered.

At present, nonparametric estimation of a joint distribution function with multivariate interval-censored data has not been considered. A current practice is to take the midpoint of the interval (L, R) as an exact observation unless it is right censored. Then Dabrowska's (1988) Kaplan-Meier estimator on the plane or van der Laan's (1996) repaired generalized maximum likelihood estimator can be applied to such data. Another practice is to treat the right endpoints of the interval-censored data as exact observations unless they are right censored (see Samuelsen and Kongerud (1994)). However, these two practices will introduce bias in the analysis (Samuelsen and Kongerud (1994)).

Multivariate right-censored data are special cases of multivariate interval-censored data. References for nonparametric estimation of distribution functions with multivariate right-censored data can be found in Campbell (1981), Hanley and Parnes (1983), Tsai *et al.* (1986), Dabrowska (1988), Gill (1992), Prentice and Cai (1992), Lin and Ying (1993), and van der Laan (1996), etc.

Nonparametric estimation of a distribution function with univariate interval-censored data has been studied by Peto (1973), Turnbull (1976), Tsai and Crowley (1985), Chang and Yang (1987), Groeneboom and Wellner (1992), Gu and Zhang (1993), and Yu *et al* (1996 and 1998), among others.

In Section 2, we discuss generalized maximum likelihood estimation of F_0 based on multivariate interval-censored data and formulate the case 2 multivariate interval censorship model. We establish consistency of the generalized maximum likelihood estimate (GMLE) of F_0 in Section 3 and asymptotic normality of the GMLE in Section 4.

2. METHOD OF ESTIMATION

Let $X = (X_1, ..., X_d)$ be a d-dimensional random survival vector with a joint distribution function $F_0(x)$, where $x = (x_1, ..., x_d)$. The observable random vector is $(L_1, R_1, ..., L_d, R_d)$, where $L_i \leq R_i$ for all i. Suppose that

$$(L_{11}, R_{11}, ..., L_{1d}, R_{1d}), ..., (L_{n1}, R_{n1}, ..., L_{nd}, R_{nd})$$

are i.i.d. copies of $(L_1, R_1, ..., L_d, R_d)$. We want to estimate the joint distribution function $F_0(x)$ (or the survival function $S_0(x) = P\{X_1 > x_1, ..., X_d > x_d\}$). Each univariate interval-censored data (L_{ij}, R_{ij}) can be viewed as an interval I_{ij}, where

$$I_{ij} = \begin{cases} [L_{ij}, R_{ij}] & \text{if} \quad L_{ij} = R_{ij}, \\ (L_{ij}, R_{ij}] & \text{if} \quad L_{ij} < R_{ij}; \end{cases}$$

therefore, each multivariate interval-censored observation can be viewed as a rectangular set $\mathcal{I}_i = I_{i1} \times \cdots \times I_{id}$, $i = 1, ..., n$.

Define a *maximal intersection* (MI), A, with respect to the $\mathcal{I}_i$'s to be a nonempty finite intersection of the $\mathcal{I}_i$'s such that for each i $A \cap \mathcal{I}_i = \varnothing$ or A. For example, let $\mathcal{I}_1 = (0, 2] \times (1, 3]$, $\mathcal{I}_2 = (0, 4] \times (1, 5]$, $\mathcal{I}_3 = (3, 5] \times (4, 8]$, and $\mathcal{I}_4 = (3, 5] \times (4, 8]$. Then the possible MI's are $(0, 2] \times (1, 3]$ and $(3, 4] \times (4, 5]$. Let $\{A_1, ..., A_m\}$ be the collection of all possible distinct MI's.

Using an argument similar to Hanley and Parnes (1983), it can be shown that the GMLE of $F_0(x)$ which maximizes the generalized likelihood function, Λ_n, must assign all the probability masses $s_1, ..., s_m$ to the sets $A_1, ..., A_m$. Thus the generalized likelihood function is as follows:

$$\Lambda_n = \prod_{i=1}^{n} \mu_F(\mathcal{I}_i) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} 1(A_j \subset \mathcal{I}_i) s_j \right], \qquad (2.1)$$

where μ_F is the measure induced by a distribution function F, $1(\cdot)$ is the indicator function, $s \ (= (s_1, ..., s_{m-1})^t) \in D_s$, $s_m = 1 - s_1 - \cdots - s_{m-1}$, s^t is the transpose of the vector s, and $D_s = \{s; s_i \geq 0, s_1 + \cdots + s_{m-1} \leq 1\}$. Denote the GMLE of s by $\hat{s}$ and that of F_0 by $\hat{F}_n$.

The $\hat{s}_j$'s can be obtained by the self-consistent algorithm described by Turnbull (1976) for univariate interval-censored data as follows: Let $s_j^{(0)} = 1/m$ for $j = 1, ..., m$. Denote $\delta_{ij} = 1(A_j \subset \mathcal{I}_i)$. At the h-step, $s_j^{(h)} = \sum_{i=1}^{n} (1/n) (\delta_{ij} s_j^{(h-1)} / \sum_{k=1}^{m} \delta_{ik} s_k^{(h-1)})$, $j = 1, ..., m$, $h \geq 1$. Repeat until the s_j's converge. The justification of the convergence of this method for multivariate interval-censored data is similar to that given in Turnbull (1976) for univariate data.

Given a GMLE $\hat{s}$, the GMLE of $F_0(\mathbf{x})$ is not uniquely defined on an MI unless the MI is a singleton. A GMLE of $F_0(\mathbf{x})$ can be obtained as follows:

$$\hat{F}_n(\mathbf{x}) = \sum_{A_j \subset [0, x_1] \times \cdots \times [0, x_d]} \hat{s}_j. \qquad (2.2)$$

Remark 1. The GMLE of s may not be unique, as the following example demonstrates.

Suppose that a sample of size 4 consists of two-dimensional interval-censored observations $(1, 6, 1, 3)$, $(1, 6, 4, 6)$, $(1, 3, 1, 6)$ and $(4, 6, 1, 6)$. Then the MIs are $A_1 = (1, 3] \times (1, 3]$, $A_2 = (1, 3] \times (4, 6]$, $A_3 = (4, 6] \times (1, 3]$ and $A_4 = (4, 6] \times (4, 6]$. $(\hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{s}_4) = r(1/2, 0, 0, 1/2) + (1 - r)(0, 1/2, 1/2, 0)$ is a GMLE of s, for all $r \in [0, 1]$. Thus there are infinitely many expressions for GMLE. However, $\mu_{\hat{F}_n}(\mathcal{I}_i) = 1/4$, $i = 1, ..., 4$, for all $r \in [0, 1]$.

In general, $\hat{s}$ may not be consistent under discrete assumptions. However, the consistency of $\hat{F}_n$ on a certain set will not be affected (for more details, see Section 3).

The derivation of the GMLE only requires that the observations $\mathcal{I}_1, ..., \mathcal{I}_n$ are i.i.d. To derive the asymptotic properties of the GMLE, we need further assumptions on F_0 and the distribution function of $(L_1, R_1, ..., L_d, R_d)$.

A set of univariate interval-censored data are referred to as *case* 2 data if they consist of strictly interval-censored, right-censored or left-censored observations, but do not contain exact observations. For such type of data, Groeneboom and Wellner (1992) formulate the case 2 univariate interval censorship model. We consider a natural multivariate extension of the case 2 univariate interval censorship model in the following.

Suppose $(U_1, V_1, ..., U_d, V_d)$ is a random censoring vector and is independent of X. The observable random vector $(L_1, R_1, ..., L_d, R_d)$ is generated by the following formula.

$$(L_i, R_i) = \begin{cases} (0, U_i) & \text{if} \quad X_i \leq U_i, \\ (U_i, V_i) & \text{if} \quad U_i < X_i \leq V_i, \quad i = 1, ..., d. \\ (V_i, +\infty) & \text{if} \quad X_i > V_i, \end{cases}$$

We call this model a *case* 2 *multivariate interval censorship model* (C2M model). In the next two sections, we shall discuss the asymptotic properties of the GMLE under the C2M model. For ease of presentation and without loss of generality (WLOG), we assumed $d = 2$ hereafter.

3. CONSISTENCY OF GMLE

In this section, we make the following assumptions under the C2M model:

$$\text{The censoring vector } (\mathbf{U}, \mathbf{V}) \text{ is discrete.} \qquad (3.1)$$

Let $\mathbf{a} = (a_1, a_2)$, $\mathbf{b} = (b_1, b_2)$, $\mathbf{U} = (U_1, U_2)$ and $\mathbf{V} = (V_1, V_2)$. Define

$$\mathscr{B} = \{(\mathbf{a}, \mathbf{b}) : g(\mathbf{a}, \mathbf{b}) > 0\}, \qquad \text{where} \quad g(\mathbf{a}, \mathbf{b}) = P(\mathbf{U} = \mathbf{a}, \mathbf{V} = \mathbf{b}),$$

Note that each point in $\mathscr{B}$ induces a grid of nine cells in R^2. Let

$$\mathscr{A}_* = \{(x_1, x_2) : x_i \in \{a_i, b_i, \pm\infty\}, i = 1, 2, (\mathbf{a}, \mathbf{b}) \in \mathscr{B}\}$$

be the set of all such grid points. We shall establish the strong consistency of the GMLE at each point in $\mathscr{A}_*$. From this we can infer the uniform strong consistency of the GMLE if F_0 is continuous and $\mathscr{A}_*$ is dense in $[0, \infty)^2$.

Let $(\mathbf{X}_i, \mathbf{U}_i, \mathbf{V}_i)$, $i = 1, ..., n$ be i.i.d. copies of $(\mathbf{X}, \mathbf{U}, \mathbf{V})$. For $(\mathbf{a}, \mathbf{b}) \in \mathscr{B}$, let

$$I_{11}(\mathbf{a}, \mathbf{b}) = (-\infty, a_1] \times (-\infty, a_2], \quad ..., \qquad\qquad ...,$$

$$I_{21}(\mathbf{a}, \mathbf{b}) = (a_1, b_1] \times (-\infty, a_2], \quad ..., \qquad\qquad ...,$$

$$I_{31}(\mathbf{a}, \mathbf{b}) = (b_1, +\infty) \times (-\infty, a_2], \quad ..., \quad I_{33}(\mathbf{a}, \mathbf{b}) = (b_1, +\infty) \times (b_2, +\infty).$$

Let $\mathscr{A}$ be the set of all vertexes of $B_1, ..., B_h$, where $B_1, ..., B_h$ are all possible MIs with respect to $I_{ij}(\mathbf{a}, \mathbf{b})$, $i, j = 1, 2, 3$, and $(\mathbf{a}, \mathbf{b}) \in \mathscr{B}$. Note that $\mathscr{A}_*$ is the set of vertexes of the rectangles $I_{ij}(\mathbf{a}, \mathbf{b})$s. Thus $\mathscr{A}_* \neq \mathscr{A}$ in general. Let

$$N_{nik}(\mathbf{a}, \mathbf{b}) = \frac{1}{n} \sum_{j=1}^{n} 1(\mathbf{X}_j \in I_{ik}(\mathbf{a}, \mathbf{b}), \mathbf{U}_j = \mathbf{a}, \mathbf{V}_j = \mathbf{b}), \qquad i, k = 1, 2, 3.$$

Then the generalized likelihood (2.1) is equal to

$$\Lambda_n(F) = \prod_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} \prod_{i=1}^{3} \prod_{j=1}^{3} [\mu_F(I_{ij}(\mathbf{a}, \mathbf{b}))]^{nN_{nij}(\mathbf{a}, \mathbf{b})},$$

where

$$\mu_F((c, d] \times (e, f]) = F(d, f) + F(c, e) - F(c, f) - F(d, e). \qquad (3.2)$$

Moreover, the normalized generalized log-likelihood function is

$$\mathscr{L}_n(F) = \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} \sum_{i=1}^{3} \sum_{j=1}^{3} N_{nij}(\mathbf{a}, \mathbf{b}) \ln[\mu_F(I_{ij}(\mathbf{a}, \mathbf{b}))].$$

Here and below we interpret $0 \log 0 = 0$ and $\log 0 = -\infty$. For this likelihood function, we let F range over the set $\mathscr{F}^*$ of all functions F on $[-\infty, +\infty]^2$ such that

$$F(+\infty, +\infty) = 1, \tag{3.3}$$

$$F(-\infty, x) = F(x, -\infty) = 0 \qquad \text{for each } x, \tag{3.4}$$

and

$$\mu_F(I) \geqslant 0 \qquad \text{for all rectangle sets } I \text{ in } (-\infty, +\infty]^2. \tag{3.5}$$

In view of (3.2), $\Lambda_n(F)$ and $\mathscr{L}_n(F)$ depend on F only through the values of F at the points $\mathbf{x} \in \mathscr{A}_*$. Because the GMLE of F_0 is not unique, we adopt expression (2.2) for the GMLE in our proofs below.

THEOREM 1.   *Under Assumption* (3.1), *the GMLE* $\hat{F}_n$ *satisfies* $\hat{F}_n(\mathbf{a}) \rightarrow F_0(\mathbf{a})$ *almost surely for all* $\mathbf{a} \in \mathscr{A}_*$.

Proof.   Verify that

$$\mathsf{L}(F) := E(\mathscr{L}_n(F)) = \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} g(\mathbf{a}, \mathbf{b}) \, h_{\mathbf{a}, \mathbf{b}}(F) \tag{3.6}$$

with

$$h_{\mathbf{a}, \mathbf{b}}(F) = \sum_{i=1}^{3} \sum_{j=1}^{3} \mu_{F_0}(I_{ij}(\mathbf{a}, \mathbf{b})) \ln[\mu_F(I_{ij}(\mathbf{a}, \mathbf{b}))].$$

Verify that the expression $h_{\mathbf{a}, \mathbf{b}}(F)$ is maximized by a function $F \in \mathscr{F}^*$ if and only if

$$\mu_F(I_{ij}(\mathbf{a}, \mathbf{b})) = \mu_{F_0}(I_{ij}(\mathbf{a}, \mathbf{b})), \qquad i, j = 1, 2, 3. \tag{3.7}$$

Equations (3.2) and (3.4) imply that (3.7) is equivalent to $F(\mathbf{x}) = F_0(\mathbf{x})$ for each vertex $\mathbf{x}$ of rectangles $I_{ij}(\mathbf{a}, \mathbf{b})$, $i, j = 1, 2, 3$. Thus F_0 maximizes $\mathsf{L}(F)$ and any other function in $\mathscr{F}^*$ that maximizes $\mathsf{L}(F)$ will coincide with F_0 on $\mathscr{A}_*$.

Note that $\mathscr{L}_n(F_0) = (1/n) \sum_{j=1}^{n} \psi(\mathbf{X}_j, \mathbf{U}_j, \mathbf{V}_j)$, where ψ is the map defined by

$$\psi(\mathbf{x}, \mathbf{a}, \mathbf{b}) = \sum_{i=1}^{3} \sum_{j=1}^{3} \mathbf{1}(\mathbf{x} \in I_{ij}(\mathbf{a}, \mathbf{b})) \ln(\mu_F(I_{ij}(\mathbf{a}, \mathbf{b}))).$$

Thus it follows from the SLLN and (3.2) that $\mathcal{L}_n(F_0) \to L(F_0)$ almost surely. By the definition of the GMLE, $\mathcal{L}_n(\hat{F}_n) \geqslant \mathcal{L}_n(F_0)$. Consequently,

$$\varliminf_{n \to \infty} \mathcal{L}_n(\hat{F}_n) \geqslant \varliminf_{n \to \infty} \mathcal{L}_n(F_0) = L(F_0) \text{ almost surely.}$$

Let Ω' denote the event on which $\varliminf_{n \to \infty} \mathcal{L}_n(\hat{F}_n) \geqslant L(F_0)$. Fix an $\omega \in \Omega'$, let $F^* \in \mathscr{F}^*$ be a limit point of $\hat{F}_{k_n}(\cdot, \omega)$ in the sense that $\hat{F}_{k_n}(\mathbf{a}, \omega) \to F^*(\mathbf{a})$ for all $\mathbf{a} \in \mathscr{A}_*$ and for some sequence $\{k_n\}$ of positive integers tending to infinity. We now show that

$$L(F^*) \geqslant L(F_0).$$

Let $t_{k_n}(\mathbf{a}, \mathbf{b})$ denote the value of the random variable $\sum_{i=1}^3 \sum_{j=1}^3 N_{k_n ij}(\mathbf{a}, \mathbf{b}) \times \ln[\mu_{\hat{F}_{k_n}}(I_{ij})]$ at the point ω. By the definition of Ω',

$$\varliminf_{n \to \infty} \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} t_{k_n}(\mathbf{a}, \mathbf{b}) \geqslant L(F_0).$$

Next, verify that

$$t_{k_n}(\mathbf{a}, \mathbf{b}) \to g(\mathbf{a}, \mathbf{b}) h_{\mathbf{a}, \mathbf{b}}(F^*)$$

for each $(\mathbf{a}, \mathbf{b}) \in \mathscr{B}$. Note also that $t_{k_n}(\mathbf{a}, \mathbf{b}) \leqslant 0$ for all $(\mathbf{a}, \mathbf{b}) \in \mathscr{B}$. From Fatou's Lemma,

$$
\begin{aligned}
\varlimsup_{n \to \infty} \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} t_{k_n}(\mathbf{a}, \mathbf{b}) &= -\varliminf_{n \to \infty} \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} -t_{k_n}(\mathbf{a}, \mathbf{b}) \\
&\leqslant -\sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} \varliminf_{n \to \infty} (-t_{k_n}(\mathbf{a}, \mathbf{b})) \\
&= \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} g(\mathbf{a}, \mathbf{b}) h_{\mathbf{a}, \mathbf{b}}(F^*) \\
&= L(F^*).
\end{aligned}
$$

Combining the above yields $L(F_0) \leqslant L(F^*)$. As F_0 maximizes L, we conclude that $L(F^*) = L(F_0)$ and therefore $F^*(\mathbf{a}) = F_0(\mathbf{a})$ for all $\mathbf{a} \in \mathscr{A}_*$. Since ω is arbitrary and Ω' has probability one, the consistency result is thus established. ∎

If $\mathscr{A}_*$ is a finite set, then it follows from the theorem that the GMLE is uniformly strongly consistent on $\mathscr{A}_*$. For arbitrary $\mathscr{A}_*$, the uniform strong consistency of the GMLE requires additional assumptions.

THEOREM 2. *Suppose that* (3.1) *holds,* F_0 *is continuous and* $\mathscr{A}_*$ *is dense in* $[0, +\infty)^2$. *Then* $\sup_{\mathbf{x} \in \mathscr{R}^2} |\hat{F}_n(\mathbf{x}) - F_0(\mathbf{x})| \to 0$ *almost surely.*

Proof. Let $F_1, F_2, ...$ be functions in $\mathscr{F}^*$ such that $F_n(\mathbf{a}) \rightarrow F_0(\mathbf{a})$ for all $\mathbf{a} \in \mathscr{A}_*$. Let M be a positive integer. Since F_0 is continuous, there is a grid which partitions the space $(-\infty, +\infty]^2$ into M disjoint rectangles $I = (c, d] \times (e, f]$ with grid points (upper-right vertexes of Is) $\mathbf{x}_1, ..., \mathbf{x}_M$ in $(-\infty, +\infty]^2$ and $\mu_{F_0}(I) \leq 1/M$ for each grid cell I. The continuity of F_0 and the fact that $\mathscr{A}_*$ is dense in $[0, +\infty)^2$ imply that there are points $\mathbf{a}_1, ..., \mathbf{a}_M$ in $\mathscr{A}_*$ such that $|F_0(\mathbf{a}_i) - F_0(\mathbf{x}_i)| \leq 1/M^2$. Using this and the facts $F_0, F_n \in \mathscr{F}^*$ and that $F_0(c, e) \leq F_0(\mathbf{x}) \leq F_0(d, f)$ and $F_n(c, e) \leq F_n(\mathbf{x}) \leq F_n(d, f)$ for each $\mathbf{x} \in I$, we derive that

$$|F_n(\mathbf{x}) - F_0(\mathbf{x})| \leq \max_{1 \leq i \leq M} |F_n(\mathbf{a}_i) - F_0(\mathbf{a}_i)| + \frac{3}{M}, \qquad \mathbf{x} \in \mathscr{R}^2.$$

This shows that F_n converges to F_0 uniformly.

By the above, the events $\bigcap_{\mathbf{a} \in \mathscr{A}_*} \{\hat{F}_n(\mathbf{a}) \rightarrow F_0(\mathbf{a})\}$ and $\{\sup_{\mathbf{x} \in \mathscr{R}^2} |\hat{F}_n(\mathbf{x}) - F_0(\mathbf{x})| \rightarrow 0\}$ are identical and thus have probability 1 by Theorem 1. ∎

Remark 2. In the case of the bivariate right censorship model, under the assumptions in Theorem 2, it is well known that the GMLE is not a consistent estimate of a continuous F_0 (see Tsai *et al.* (1986)).

4. ASYMPTOTIC NORMALITY OF GMLE

Under the univariate case 2 interval censorship model, Groeneboom and Wellner (1992) conjecture that if the censoring distribution is continuous, then the GMLE of a continuous F_0 is not asymptotically normally distributed and the convergence rate is not in $\sqrt{n}$. Yu *et al.* (1998) prove that if the censoring vector takes on finitely many values, then under an additional assumption the GMLE is asymptotically normally distributed and the convergence rate is in $\sqrt{n}$. In the multivariate case, the situation is more complicated. In this section we shall obtain the asymptotic normality of the GMLE under the C2M model and the assumptions that

$$\mathscr{A}_* \text{ contains finitely many elements,} \qquad (4.1)$$

$$\mu_{F_0}((a_1, b_1] \times (a_2, b_2]) > 0 \quad \text{if} \quad \mathbf{a}, \mathbf{b} \in \mathscr{A}_* \quad \text{and} \quad a_i < b_i, \quad i = 1, 2. \quad (4.2)$$

and

$$\mathscr{A}_* = \mathscr{A} \quad \text{(see Section 3)}. \qquad (4.3)$$

Note that under the current assumptions the standard method for finite parametric models can be used.

Remark 3. The GMLE of s may not be unique (see Remark 1) and Theorem 1 does not ensure the consistency of the GMLE $\hat{s}$ as $\mathscr{A}$ and $\mathscr{A}_*$ are not the same in general. Note that the consistency of the GMLE $\hat{F}_n$ on $\mathscr{A}_*$ is mainly due to Eq. (3.7), since $\mathscr{A}_*$ is the set of all vertexes of the rectangles $I_{ij}(\mathbf{a}, \mathbf{b})$'s.

By Theorem 1 and (4.3), the GMLE $\hat{F}_n$ is consistent on the set $\mathscr{A}$. Since $\hat{s}_j = \mu_{\hat{F}_n}(A_j)$, where the vertexes of the MI A_j belong to $\mathscr{A}$, $\hat{s}$ is consistent by (3.2).

Let $s_j^o = \mu_{F_0}(A_j)$. Then (4.2) yields $s_j^o > 0$ for all j. Verify that (3.6) yields

$$
\begin{aligned}
\mathŁ(F) &= \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} g(\mathbf{a}, \mathbf{b}) \sum_{i=1}^{3} \sum_{l=1}^{3} \sum_k s_k^o \mathbf{1}(A_k \subset I_{il}(\mathbf{a}, \mathbf{b})) \\
&\quad \times \ln \sum_j s_j \mathbf{1}(A_j \subset I_{il}(\mathbf{a},)) \\
&= \sum_{(\mathbf{a}, \mathbf{b}) \in \mathscr{B}} \sum_{i=1}^{3} \sum_{l=1}^{3} \left[g(\mathbf{a}, \mathbf{b}) \sum_k s_k^o \mathbf{1}(A_k \subset I_{il}(\mathbf{a}, \mathbf{b})) \right] \\
&\quad \times \ln \sum_j s_j \mathbf{1}(A_j \subset I_{il}(\mathbf{a}, \mathbf{b})).
\end{aligned}
\tag{4.4}
$$

Let

$$
\{I_1, ..., I_\beta\} = \{I_{ij}(\mathbf{a}, \mathbf{b}): i, j = 1, 2, 3, (\mathbf{a}, \mathbf{b}) \in \mathscr{B}\},
$$

and

$$
p_h = g(\mathbf{a}, \mathbf{b}) \sum_k s_k^o \mathbf{1}(A_k \subset I_{il}(\mathbf{a}, \mathbf{b})).
$$

We can rewrite (4.4) as

$$
\mathŁ(F) = \sum_{h=1}^{\beta} p_h \ln \sum_{j=1}^{m} s_j \mathbf{1}(A_j \subset I_h) = \sum_{h=1}^{\beta} p_h \ln \sum_{j=1}^{m} s_j \delta_{hj}.
$$

From (4.2), $p_h > 0$, $h = 1, ..., \beta$. Set $J = -E(\partial^2 \mathscr{L}(F_0)/\partial s \, \partial s')$, where $\partial \mathscr{L}/\partial s$ is an $(m-1) \times 1$ vector and $\partial^2 \mathscr{L}/\partial s \, \partial s'$ is an $(m-1) \times (m-1)$ matrix. Verify that

$$J = nE\left(\frac{\partial \mathscr{L}(F_0)}{\partial \mathbf{s}} \frac{\partial \mathscr{L}(F_0)}{\partial \mathbf{s}^t}\right) = -\frac{\partial^2 \mathbf{L}}{\partial \mathbf{s}\, \partial \mathbf{s}^t}$$

$$= \left(\sum_{h=1}^{\beta} p_h \frac{(\delta_{hi} - \delta_{hm})(\delta_{hj} - \delta_{hm})}{(\sum_{k=1}^{m} \delta_{hk} s_k^o)^2}\right)_{(m-1)\times(m-1)} = UU^t,$$

where

$$U = \begin{pmatrix} \dfrac{(\delta_{11} - \delta_{1m})\sqrt{p_1}}{\sum_{k=1}^{m} \delta_{1k} s_k^o} & \cdots & \dfrac{(\delta_{\beta 1} - \delta_{\beta m})\sqrt{p_\beta}}{\sum_{k=1}^{m} \delta_{\beta k} s_k^o} \\ \cdot & \cdots & \cdot \\ \dfrac{(\delta_{1(m-1)} - \delta_{1m})\sqrt{p_1}}{\sum_{k=1}^{m} \delta_{1k} s_k^o} & \cdots & \dfrac{(\delta_{\beta(m-1)} - \delta_{\beta m})\sqrt{p_\beta}}{\sum_{k=1}^{m} \delta_{\beta k} s_k^o} \end{pmatrix}.$$

We now show that J is nonsingular. Let $\mathbf{x}_j$ be the upper-right vertex of A_j, $j = 1, ..., m - 1$. By reordering the I_j's, WLOG, we can assume that the upper-right vertex of I_i is equal to $\mathbf{x}_i$, $i = 1, ..., m - 1$. Thus $I_i \cap A_j = \varnothing$ for $j > i$, $i = 1, ..., m - 1$. Then the matrix U has the upper triangle matrix from

$$U = \begin{pmatrix} \dfrac{\sqrt{p_1}}{s_1^o} & \cdot & \cdots & \cdot & \cdots & \dfrac{(\delta_{\beta 1} - \delta_{\beta m})\sqrt{p_\beta}}{\sum_{k=1}^{m} \delta_{\beta k} s_k^o} \\ 0 & \dfrac{\sqrt{p_2}}{s_2^o + \delta_{21} s_1^o} & \cdots & \cdot & \cdots & \dfrac{(\delta_{\beta 2} - \delta_{\beta m})\sqrt{p_\beta}}{\sum_{k=1}^{m} \delta_{\beta k} s_k^o} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \dfrac{\sqrt{p_{m-1}}}{s_{m-1}^o + \sum_{k=1}^{m-2} \delta_{(m-1)k} s_k^o} & \cdots & \dfrac{(\delta_{\beta(m-1)} - \delta_{\beta m})\sqrt{p_\beta}}{\sum_{k=1}^{m} \delta_{\beta k} s_k^o} \end{pmatrix}.$$

Recall $s_i^o > 0$ and $p_i > 0$ for $i = 1, ..., m - 1$. It follows that the matrix U is of full rank and $J = UU^t$ is nonsingular.

It is easy to verify that

$$\frac{\partial^2 \mathscr{L}(\hat{F}_n)}{\partial \mathbf{s}\, \partial \mathbf{s}^t} \to E\left(\frac{\partial^2 \mathscr{L}(F_0)}{\partial \mathbf{s}\, \partial \mathbf{s}^t}\right) = -J.$$

It thus follows that

$$\frac{\partial \mathscr{L}(\hat{F}_n)}{\partial \mathbf{s}} = \frac{\partial \mathscr{L}(F_0)}{\partial \mathbf{s}} - J\Delta_n + o_p(\|\Delta_n\|),$$

where Δ_n is the $(m-1)$-dimensional column vector with entries $\hat{s}_i - s_i^o = \mu_{\hat{F}_n}(A_i) - \mu_{F_0}(A_i)$, $i = 1, ..., m-1$. Let $\Omega_n = \{\inf_{i \leqslant m} \hat{s}_i = 0\}$. Verify that

$$0 = \frac{\partial \mathscr{L}(\hat{F}_n)}{\partial \mathbf{s}} \text{ except on the event } \Omega_n,$$

and by Theorem 1 and Assumptions (4.1) and (4.2),

$$P(\Omega_n) \to 0 \qquad \text{as} \qquad n \to \infty.$$

It follows from the CLT that $\sqrt{n}\,(\partial \mathscr{L}(F_0)/\partial \mathbf{s})$ is asymptotically normal with mean 0 and dispersion matrix J. This shows that $\Delta_n = J^{-1} \times (\partial \mathscr{L}(F_0)/\partial \mathbf{s}) + o_p(n^{-1/2})$. Thus we have the following result.

THEOREM 3. *Under Assumptions (4.1), (4.2) and (4.3),*

$$\sqrt{n} \begin{pmatrix} \hat{s}_1 - s_1^o \\ \vdots \\ \hat{s}_{m-1} - s_{m-1}^o \end{pmatrix}$$

is asymptotically normal with mean 0 and dispersion matrix J^{-1}. A strongly consistent estimator of J is given by $\hat{J} = -(\partial^2 \mathscr{L}(\hat{F}_n)/\partial \mathbf{s}\, \partial \mathbf{s}^t)$. Furthermore, $\sqrt{n}\, [\hat{F}_n(\mathbf{x}) - F_0(\mathbf{x})]$ is asymptotically normally distributed for all $\mathbf{x} \in \mathscr{A}_*$. A consistent estimate of the asymptotic variance of $\hat{F}_n(\mathbf{x})$ is $(1/n)\, \mathbf{c}^t \hat{J}^{-1} \mathbf{c}$, where $\mathbf{c}$ is a $(m-1) \times 1$ vector with the ith entry $c_i = 1(A_i \subset [0, x_1] \times [0, x_2])$ unless $F_0(\mathbf{x}) = 1$.

Under the assumptions in Theorem 3, the GMLE is also asymptotically efficient. The proof of this assertion is straightforward and is omitted.

ACKNOWLEDGMENT

REFERENCES

G. Campbell, Nonparametric bivariate estimation with randomly censored data, *Biometrika* **68** (1981), 417–422.

M. N. Chang and G. Yang, Strong consistency of a self-consistent estimator of the survival function with doubly censored data, *Ann. Statist.* **15** (1987), 1536–1547.

D. G. Clayton, A model of association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika* **65** (1978), 141–151.

D. G. Clayton and J. Cuzick, Multivariate generalizations of the proportional hazards model (with discussion), *J. Roy. Statist. Soc. Ser. A* **148** (1985), 82–117.

D. M. Dabrowska, Kaplan-Meier estimate on the plane, *Ann. Statist.* **16** (1988), 1475–1489.

R. Gill, Multivariate survival analysis, *Theory Prob. Appl.* **37** (1992), 18–31.

P. Groeneboom and J. A. Wellner, "Information Bounds and Nonparametric Maximum Likelihood Estimation," Birkhäuser, Basel, 1992.

M. G. Gu and C. H. Zhang, Asymptotic properties of self-consistent estimator based on doubly censored data, *Ann. Statist.* **21** (1993), 611–624.

J. A. Hanley and M. N. Parnes, Nonparametric estimation of a multivariate distribution in the presence of censoring, *Biometrics* **39** (1983), 129–139.

The Italian-American Cataract Study Group, Incidence and progression of cortical, nuclear, and posterior subcapsular cataracts, *Amer. J. Ophthal.* **118** (1994), 623–631.

E. L. Kaplan and P. Meier, Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.* **53** (1958), 457–481.

D. Y. Lin and Z. L. Ying, A simple nonparametric estimator of the bivariate survival function under univariate censoring, *Biometrika* **80** (1993), 573–581.

R. Peto, Experimental survival curves for interval-censored data, *Appl. Statist.* **22** (1973), 86–91.

R. L. Prentice and J. Cai, Covariance and survival function estimation using censored multivariate failure time data, *Biometrika* **79**(3) (1992), 495–512.

S. O. Samuelsen and J. Kongerud, Interval censoring in longitudinal data of respiratory symptoms in aluminum potroom workers: A comparison of methods, *Statist. Medicine* **13** (1994), 1771–1780.

W. Y. Tsai and J. Crowley, A large sample study of the generalized maximum likelihood estimators from incomplete data via self-consistency, *Ann. Statist.* **13** (1985), 1317–1334.

W. Y. Tsai, S. Leurgans, and J. Crowley, Nonparametric estimation of a bivariate survival function in the presence of censoring, *Ann. Statist.* **14** (1986), 1351–1365.

B. W. Turnbull, The empirical distribution function with arbitrary grouped, censored and truncated data, *J. Roy. Statist. Soc. Ser. B* **38** (1976), 290–295.

M. J. Van der Laan, Efficient estimation in the bivariate censoring model and repairing NPMLE, *Ann. Statist.* **24** (1996), 596–627.

Q. Q. Yu, A. Schick, L. L. Li, and G. Y. C. Wong, Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times, *Canad. J. Statist.*, to appear.

Q. Q. Yu, A. Schick, L. L. Li, and G. Y. C. Wong, Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data, *Statist. Prob. Lett.* **37** (1998), 223–228.

CONSISTENCY OF THE GMLE
WITH MIXED CASE INTERVAL-CENSORED DATA

By Anton Schick and Qiqing Yu

Binghamton University

Abstract. In this paper we consider an interval censorship model in which the endpoints of the censoring intervals are determined by a two stage experiment. In the first stage the value k of a random integer is selected; in the second stage the endpoints are determined by a case k interval censorship model. We prove the strong consistency in the $L_1(\mu)$-topology of the nonparametric maximum likelihood estimate of the underlying survival function for a measure μ which is derived from the distributions of the endpoints. This consistency result yields strong consistency for the topologies of weak convergence, pointwise convergence and uniform convergence under additional assumptions. These results improve and generalize existing ones in the literature.

1. Introduction

In industrial life testing and medical research, one is frequently unable to observe the random variable X of interest directly, but can observe a pair (L, R) of extended random variables such that

$$-\infty \leq L < X \leq R \leq \infty.$$

For example consider an animal study in which a mouse has to be dissected to check whether a tumor has developed. At the time of dissection we can only infer whether the tumor is present, or has not yet developed. Thus, if we let X denote the onset of tumor and Y the time of the dissection, then the corresponding pair (L, R) is given by

$$(L, R) = \begin{cases} (-\infty, Y), & X \leq Y, \\ (Y, \infty), & X > Y. \end{cases}$$

If X and Y are independent, then this model is called the case 1 interval censorship model (Groeneboom and Wellner (1992)) and the data pair (L, R) is usually replaced by the current status data

1

$(Y, I[X \leq Y])$, where $I[A]$ is the indicator function of the set A. Examples of the current status data are mentioned in Ayer et al. (1955), Keiding (1991) and Wang and Gardiner (1996).

Another interval censorship model is the case 2 model considered by Groeneboom and Wellner (1992). Consider an experiment with two inspection times U and V such that $U < V$ and (U, V) is independent of X. One can only determine whether X occurs before time U, between times U and V or after time V. More formally, one observes the random vector $(U, V, I[X \leq U], I[U < X \leq V])$. In this model

$$
(L, R) = \begin{cases} (-\infty, U), & X \leq U, \\ (U, V), & U < X \leq V, \\ (V, \infty), & X > V. \end{cases}
$$

Note that (L, R) is a function of the random vector $(U, V, I[X \leq U], I[U < X \leq V])$. However, V cannot be recovered from the pair (L, R) on the event $\{X \leq U\}$. Thus the pair (L, R) carries less information than the vector $(U, V, I[X \leq U], I[U < X \leq V])$.

The case 1 and case 2 models are special cases of the case k model (Wellner, 1995). In this model there are k inspection times $Y_1 < \cdots < Y_k$ which are independent of X, and one observes into which of the random intervals $(-\infty, Y_1], \ldots, (Y_k, \infty)$ the random variable X belongs. Note that the case k model for $k > 2$ can be formally reduced to a case 2 model with U and V functions of X and the inspection times $Y_1, \ldots, Y_k$. The resulting U and V are then no longer independent of X violating a key assumption used in deriving consistency results for the case 2 model.

While the case 1 model gives a good description of the animal study mentioned above, a data set from a case k model ($k \geq 2$) is difficult to find in medical research since it is very unlikely that every patient under study has exactly the same number of visits. Finkelstein and Wolfe (1985) presented a closely related type of interval-censored data in comparing two different treatments for breast cancer patients. The censoring intervals arose in the follow-up studies for patients treated with radiotherapy and chemotherapy. The failure time X is the time until cosmetic deterioration as determined by the appearance of breast retraction. Each patient had several follow-ups and the number of follow-ups differed from patient to patient. One only knows that the failure time occurred either before the first follow-up, or after the last follow-up or between two consecutive follow-ups. Other examples of such type of interval-censored data can be found in AIDS studies (Becker and Melbye (1991); Aragon and Eberly (1992)).

In this paper we assume that the pair (L, R) is generated as a mixture of case k models. This formulation encompasses the various case k models and the data setting occurring in Finkelstein and Wolfe (1985). A precise definition of this mixture model is given in Section 2.

Let F_0 denote the unknown distribution function of X. This distribution function is commonly estimated by the generalized maximum likelihood estimate (GMLE). Ayer et al. (1955) derived an explicit expression of the GMLE for the case 1 model. However, in general the GMLE does not have an explicit solution. In deriving a numerical solution for the GMLE, Peto (1973) used the Newton-Raphson algorithm; Turnbull (1976) proposed a self-consistent algorithm; Groeneboom and Wellner (1992) proposed an iterative convex minorant algorithm. A detailed discussion of

some computational aspects is given in Wellner and Zhan (1997).

Various consistency results are available for the GMLE. In the case 1 model, Ayer et al. (1955) proved the weak consistency of the GMLE at continuity points of F_0 under additional assumptions on G, the distribution function of Y. The uniform strong consistency of the GMLE has been established by Groeneboom and Wellner (1992), van de Geer (1993, Example 3.3a), Wang and Gardiner (1996) and Yu et al. (1998a) for continuous F_0 using various assumptions and techniques. In the case 2 model, the uniform strong consistency of the GMLE has been established by Groeneboom and Wellner (1992), van de Geer (1993, Example 3.3b), and Yu et al. (1998b) for continuous F_0.

In Section 2 we shall obtain the strong $L_1(\mu)$-consistency of the GMLE for our mixture of case k models for some measure μ. This result shows that the $L_1(\mu)$-topology is the appropriate topology as it gives consistency without additional assumptions in the case k models. Convergence in stronger topologies such as the topologies of weak convergence and uniform convergence requires additional conditions. This is pursued in Section 3. In the process we also point out some erroneous consistency claims in the literature. The proof of the $L_1(\mu)$-consistency is given in Section 4. It exploits the special structure of the likelihood for this model and does not require any advanced theory. Section 5 collects various other proofs.

2. Main Results

We begin by giving a precise definition of our model. This is done by describing how the endpoints L and R are generated. Let K be a positive random integer and $\mathbf{Y} = \{Y_{k,j} : k = 1, 2, \ldots, j = 1, \ldots, k\}$ be an array of random variables such that $Y_{k,1} < \cdots < Y_{k,k}$. Assume throughout that $(K, \mathbf{Y})$ and X are independent. On the event $\{K = k\}$, let (L, R) denote the endpoints of that random interval among $(-\infty, Y_{k,1}], (Y_{k,1}, Y_{k,2}], \ldots, (Y_{k,k}, \infty)$ which contains X. We refer to this model as the *mixed case* model as it can be viewed as a mixture of the various case k models.

In some clinical studies, an examination is performed at the start of the study and follow-ups are scheduled one at a time till the end of the study. This can be modeled by taking $Y_{k,j} = \sum_{i=1}^{j-1} \xi_i$ and $K = \sup\{k \geq 1 : \sum_{i=1}^{k-1} \xi_i \leq \tau\}$, where $\xi_1, \xi_2, \ldots$ denote the (positive) inter-follow-up times and τ is the length of the study. In this case K may not be bounded. For example, if the inter-follow-up times are independent with a common exponential distribution, then $K - 1$ is a Poisson random variable; thus K is unbounded, yet $E(K) < \infty$. In general, if the inter-follow-up times are independent and identically distributed, then $E(K) < \infty$.

To define the GMLE, let $(L_1, R_1), \ldots, (L_n, R_n)$ be independent copies of the pair (L, R) defined above and define the generalized likelihood function Λ_n by

$$\Lambda_n(F) = \prod_{j=1}^{n} [F(R_j) - F(L_j)], \quad F \in \mathcal{F},$$

3

where $\mathcal{F}$ is the collection of all nondecreasing functions F from $[-\infty, +\infty]$ into $[0, 1]$ with $F(-\infty) = 0$ and $F(+\infty) = 1$. We think of F_0 as a member of $\mathcal{F}$. Note that $\Lambda_n(F)$ depends on F only through the values of F at the points L_j or R_j, $j = 1, \ldots, n$. Thus there exists no unique maximizer of $\Lambda_n(F)$ over the set $\mathcal{F}$. However, there exists a unique maximizer $\hat{F}_n$ over the set $\mathcal{F}$ which is right continuous and piecewise constant with possible discontinuities only at the observed values of L_j and R_j, $j = 1, \ldots, n$. We call this maximizer $\hat{F}_n$ the GMLE of F_0.

Define a measure μ on the Borel σ-field $\mathcal{B}$ on $\mathbb{R}$ by

$$\mu(B) = \sum_{k=1}^{\infty} P(K = k) \sum_{j=1}^{k} P(Y_{k,j} \in B \mid K = k), \qquad B \in \mathcal{B}.$$

We are now ready to state our main result, namely the (strong) $L_1(\mu)$ consistency of the GMLE.

2.1. Theorem. *Let $E(K) < \infty$. Then $\int |\hat{F}_n - F_0| \, d\mu \to 0$ almost surely.*

The condition $E(K) < \infty$ implies the finiteness of the measure μ and of the expectation $E[\log(F_0(R) - F_0(L))]$. These two latter conditions play an important role in our proof given in Section 4.

One referee pointed out that results of van de Geer's (1993) (namely her Lemma 1.1 and Theorem 3.1) may be used to prove a result very similar to our Theorem 2.1 with the help of some inequalities suggested by this referee. This alternative proof leads to $L_1(\tilde{\mu})$-consistency for some finite measure $\tilde{\mu}$ that is equivalent to our measure μ and does not require the finiteness of $E(K)$. Actually, such a result implies our result in view of the following simple lemma which we state without a proof.

2.2. Lemma. *Let μ_1 and μ_2 be two finite measures and $g, g_1, g_2, \ldots$ be measurable functions into $[0, 1]$. Suppose that μ_2 is absolutely continuous with respect to μ_1. Then $\int |g_n - g| \, d\mu_1 \to 0$ implies $\int |g_n - g| \, d\mu_2 \to 0$.*

We have decided to present our original proof since it is direct and elementary and since $E(K) < \infty$ is a rather mild assumption that is typically satisfied in applications.

In the remainder of this section we mention some corollaries of Theorem 2.1. The first one is of interest when the inspection times are discrete. It follows from the fact that $\mu(\{a\})|\hat{F}_n(a) - F_0(a)| \le \int |\hat{F}_n - F_0| \, d\mu$ for every $a \in \mathbb{R}$ and generalizes the consistency results given in Yu et al. (1998a,b) for the case 1 and case 2 models with discrete inspection times.

2.3. Corollary. *Let $E(K) < \infty$. Then $\hat{F}_n(a) \to F_0(a)$ almost surely for each point a with $\mu(\{a\}) > 0$.*

In the next corollary we state results for a measure ν that depends on the distribution of L and R and is easier to interpret than μ. We take ν to be the sum of the marginal distributions of L and R:

$$\nu(B) = P(L \in B) + P(R \in B), \quad B \in \mathcal{B}.$$

In view of the set inclusion

$$\{L \in B\} \cup \{R \in B\} \subset \bigcup_{k=1}^{\infty} \bigcup_{i=1}^{k} \{K = k, Y_{k,i} \in B\},$$

we have $\nu(B) \leq 2\mu(B)$. Thus we immediately get the following corollary.

2.4. Corollary. *Let $E(K) < \infty$. Then the following are true.*

(1) $\int |\hat{F}_n - F_0| \, d\nu \to 0$ *almost surely.*

(2) $\hat{F}_n(a) \to F_0(a)$ *almost surely for each point a with $\nu(\{a\}) > 0$.*

3. Other Consistency Results

In this section we shall show that under additional assumptions strong $L_1(\mu)$-consistency implies strong consistency in other topologies such as the topologies of weak convergence, pointwise convergence and uniform convergence. Throughout we always assume that $E(K)$ is finite so that μ is a finite measure and $P(\Omega_\mu) = 1$ by Theorem 2.1, where

$$\Omega_\mu = \{ \lim_{n \to \infty} \int |\hat{F}_n - F_0| \, d\mu = 0 \}.$$

Although the results of this section are formulated for the measure μ defined in the previous section, they are true for any finite measure for which the GMLE is strongly L_1-consistent as only the finiteness of μ and $P(\Omega_\mu) = 1$ are used in their proofs. These proofs are deferred to Section 5.

Let a be a real number. We call a a *support point of μ* if $\mu((a - \epsilon, a + \epsilon)) > 0$ for every $\epsilon > 0$. We call a *regular* if $\mu((a - \epsilon, a]) > 0$ and $\mu([a, a + \epsilon)) > 0$ for all $\epsilon > 0$. We call a *strongly regular* if $\mu((a - \epsilon, a)) > 0$ and $\mu([a, a + \epsilon)) > 0$ for all $\epsilon > 0$. We call a a *point of increase of F_0* if $F_0(a + \epsilon) - F_0(a - \epsilon) > 0$ for each $\epsilon > 0$.

In view of the inequality $\nu \leq 2\mu$, sufficient conditions for the first three of the above concepts are obtained by replacing μ be ν. As these sufficient conditions are in terms of the distribution of L and R, they are easier to interpret and thus more meaningful from an applied point of view.

Ayer et al. (1955) established the weak consistency of the GMLE at regular continuity points of F_0 in the case 1 model. Our first proposition gives a strong consistency result for regular continuity points in our more general model.

3.1. Proposition. *For each $\omega \in \Omega_\mu$ and each regular continuity point a of F_0, $\hat{F}_n(a; \omega) \to F_0(a)$.*

The next two propositions address weak convergence on an open interval and on the entire line.

3.2. Proposition. *Suppose every point in an open interval (a, b) is a support point of μ. Then $\hat{F}_n(x; \omega) \to F_0(x)$ for every continuity point x of F_0 in (a, b) and every $\omega \in \Omega_\mu$. If also $F_0(a) = 0$ and $F_0(b-) = 1$, then $\hat{F}_n(x; \omega) \to F_0(x)$ for all continuity points x of F_0 and all $\omega \in \Omega_\mu$.*

5

3.3. Proposition. *If every point of increase of F_0 is strongly regular, then $\hat{F}_n(x; \omega) \to F_0(x)$ for all continuity points of F_0 and all $\omega \in \Omega_\mu$.*

Combining these propositions with Corollary 2.3 yields the following results on pointwise convergence on open intervals and on the entire line.

3.4. Corollary. *Suppose every point x in an open interval (a, b) is a support point of μ and satisfies $\mu(\{x\}) > 0$ if x is a discontinuity point of F_0. Then $\hat{F}_n(x; \omega) \to F_0(x)$ for every x in (a, b) and every $\omega \in \Omega_\mu$. Moreover, if $F_0(a) = 0$ and $F_0(b-) = 1$, then $\hat{F}_n(x; \omega) \to F_0(x)$ for all $x \in \mathbb{R}$ and all $\omega \in \Omega_\mu$.*

3.5. Corollary. *If every point of increase of F_0 is strongly regular and if $\mu(\{a\}) > 0$ for each discontinuity point a of F_0, then $\hat{F}_n(x; \omega) \to F_0(x)$ for all $x \in \mathbb{R}$ and all $\omega \in \Omega_\mu$.*

The next proposition addresses uniform convergence.

3.6. Proposition. *Suppose that F_0 is continuous and that, for all $a < b$, $0 < F_0(a) < F_0(b) < 1$ implies $\mu((a, b)) > 0$. Then the GMLE is uniformly strongly consistent, i.e.,*

$$\sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)| \to 0 \quad a.s..$$

This proposition generalizes the strong uniform consistency results given by Groeneboom and Wellner (1992) for the case 1 and 2 models. In the case 1 model they require that F_0 and G, the distribution function of Y, are continuous and that the probability measure μ_{F_0} induced by F_0 is absolutely continuous with respect to μ ($\mu_{F_0} << \mu$). Proposition 3.6 does not require the continuity of G and weakens the absolute continuity requirement. In the case 2 model Groeneboom and Wellner assume that F_0 is continuous and that the joint distribution of U and V has a Lebesgue density g such that $g(u, v) > 0$ if $0 < F_0(u) < F_0(v) < 1$. Their assumption implies that the measure μ has a Lebesgue density which is positive on the set $\{t : 0 < F_0(t) < 1\}$ and therefore implies that $\mu((a, b)) > 0$ if $0 < F_0(a) < F_0(b) < 1$. Consequently, Proposition 3.6 improves and generalizes their result.

Proposition 3.6 also generalizes the strong uniform consistency results given by van de Geer (1993) for the case 1 and 2 models under the assumption that F_0 is continuous and $\mu_{F_0} << \mu$. The latter implies that $\mu((a, b)) > 0$ if $0 < F_0(a) < F_0(b) < 1$. However, if μ is discrete, its support is dense in $(0, +\infty)$, and F_0 is exponential, then the assumption in Proposition 3.6 is satisfied, but $\mu_{F_0} << \mu$ is not true.

In clinical follow-ups, the studies typically last for a certain period of time, say $[\tau_1, \tau_2]$. It is often that $F_0(\tau_2) < 1$ in which case the conditions in Proposition 3.6 are not satisfied. In this regard, Gentleman and Geyer (1994) claimed a vague convergence result in their Theorem 2 and Huang (1996) claimed a uniform strong consistency result in his Theorem 3.1. Both of their results as stated imply the uniform strong consistency of the GMLE on $[\tau_1, \tau_2]$ in the case 1 model, if F_0

is continuous and the inspection time Y is uniformly distributed on $[\tau_1, \tau_2]$. The following example shows that this is not true.

3.7. Example. Consider current status data $(Y_1, I[X_1 \le Y_1]), \ldots, (Y_n, I[X_n \le Y_n])$, where the survival times $X_1, \ldots, X_n$ are uniformly distributed on $[0, 3]$ and the inspection times $Y_1, \ldots, Y_n$ are uniformly distributed on $[1, 2]$. Then F_0 is the uniform distribution function on $[0, 3]$ and μ is the uniform distribution on $[1, 2]$. Note that on the event $\bigcup_{j=1}^{n}\{X_j > 2 > Y_j, Y_j < Y_i, i = 1, \ldots, n, i \ne j\}$ we have $\hat{F}_n(1) = 0$, and on the event $\bigcup_{j=1}^{n}\{X_j \le 1 \le Y_j, Y_j > Y_i, i = 1, \ldots, n, i \ne j\}$ we have $\hat{F}_n(2) = \hat{F}_n(2-) = 1$. Both events have probability $1/3$. Since $F_0(1) = 1/3$ and $F_0(2) = F_0(2-) = 2/3$, we see that $\hat{F}_n(x)$ does not converge to $F_0(x)$ almost surely for $x = 1, 2$ and $\hat{F}_n(2-)$ does not converge to $F_0(2-)$ almost surely. This shows that pointwise convergence on the closed interval $[\tau_1, \tau_2]$ to a continuous F_0 is not implied by the condition: $\mu([a, b]) > 0$ for all a and b such that $\tau_1 \le a < b \le \tau_2$.

The following proposition indicates how to fix the assumptions.

3.8. Proposition. *Suppose the following four conditions hold for real numbers $\tau_1 < \tau_2$.*

(1) *F_0 is continuous at every point in the interval $(\tau_1, \tau_2]$;*
(2) *either $\mu(\{\tau_1\}) > 0$ or $F_0(\tau_1) = 0$;*
(3) *either $\mu(\{\tau_2\}) > 0$ or $F_0(\tau_2-) = 1$;*
(4) *for all a and b in (τ_1, τ_2), $0 < F_0(a) < F_0(b) < 1$ implies $\mu((a, b)) > 0$.*

Then the GMLE is uniformly strongly consistent on $[\tau_1, \tau_2]$, i.e.,

$$\sup_{x \in [\tau_1, \tau_2]} |\hat{F}_n(x) - F_0(x)| \to 0 \quad a.s..$$

4. Proof of Theorem 2.1

Recall that L may take the value $-\infty$ and R the value $+\infty$. The normalized log-likelihood is

$$\mathcal{L}_n(F) = \frac{1}{n} \sum_{j=1}^{n} \log [F(R_j) - F(L_j)], \quad F \in \mathcal{F}.$$

By the strong law of large numbers (SLLN), $\mathcal{L}_n(F)$ converges almost surely to its mean

$$\mathcal{L}(F) = E(\log [F(R) - F(L)]) = \sum_{k=1}^{\infty} P(K = k) E(h_{F,k}(Y_{k,1}, \ldots, Y_{k,k}) \mid K = k),$$

where

$$h_{F,k}(y_1, \ldots, y_k) = \sum_{j=0}^{k} (F_0(y_{j+1}) - F_0(y_j)) \log(F(y_{j+1}) - F(y_j)),$$

7

for $-\infty = y_0 < y_1 < \cdots < y_k < y_{k+1} = \infty$. Here and below we interpret $0 \log 0 = 0$ and $\log 0 = -\infty$.

It is easy to check that, for each positive integer k and real numbers $y_1 < \cdots < y_k$, the expression $h_{F,k}(y_1, \ldots, y_k)$ is maximized by a function $F \in \mathcal{F}$ if and only if $F(y_j) = F_0(y_j)$ for $j = 1, \ldots, k$. Since $\sup\{|p \log p| : 0 \leq p \leq 1\} < 1$, $|h_{F_0,k}|$ is bounded by k. Since K has finite expectation, we see that $\mathcal{L}(F_0)$ is finite. Hence F_0 maximizes $\mathcal{L}(\cdot)$ over the set $\mathcal{F}$ and any other function $F \in \mathcal{F}$ that maximizes $\mathcal{L}(\cdot)$ satisfies that $F = F_0$ a.e. μ.

Let $\{F_n\}$ be a sequence in $\mathcal{F}$. By a pointwise limit of this sequence we mean an $F \in \mathcal{F}$ such that $F_{n'}(x) \to F(x)$ for all $x \in \mathbb{R}$ and some subsequence $\{n'\}$. Helly's selection theorem (Rudin (1976), pg 167) guarantees the existence of pointwise limits. Let now Ω' be the set of all sample points ω for which the sequence $\{\hat{F}_n(\cdot; \omega)\}$ has only pointwise limits F such that $\mathcal{L}(F) \geq \mathcal{L}(F_0)$. In view of the above discussion, for each $\omega \in \Omega'$, all the limit points of $\{\hat{F}_n(\cdot; \omega)\}$ equal F_0 a.e. μ and this gives that $\int |\hat{F}_n(x; \omega) - F_0(x)| \, d\mu(x) \to 0$. Thus the desired result follows if we show that Ω' has probability 1. Let $\hat{Q}_n$ denote the empirical estimator of Q, the distribution of (L, R). By the SLLN, $\Omega_0 = \{\mathcal{L}_n(F_0) \to \mathcal{L}(F_0)\}$ has probability 1, and so does $\Omega_U = \{\hat{Q}_n(U) \to Q(U)\}$ for every Borel subset U of $\Delta = \{(l, r) : -\infty \leq l < r \leq \infty\}$. Thus we are done if we show that Ω' contains the intersection Ω_* of Ω_0 and $\bigcap_{U \in \mathcal{U}} \Omega_U$ for some countable collection $\mathcal{U}$ of Borel subsets of Δ.

Let α be a positive integer. Then there are finitely many extended real numbers

$$-\infty = q_0 < q_1 < q_2 < \cdots < q_\beta = \infty$$

such that $\mu((q_{i-1}, q_i)) < 2^{-\alpha}$ for $i = 1, \ldots, \beta$. Now form the sets $U_0, \ldots, U_{2\beta}$ by setting $U_{2i-1} = (q_{i-1}, q_i)$ for $i = 1, \ldots, \beta$, and $U_{2i} = [q_i, q_i]$ for $i = 0, \ldots, \beta$. Let $\mathcal{U}_\alpha$ denote the collection of all nonempty sets of the form $U_{ij} = \Delta \cap (U_i \times U_j)$ for $0 \leq i \leq j \leq 2\beta$. We shall take $\mathcal{U} = \bigcup_\alpha \mathcal{U}_\alpha$.

Let now ω belong to Ω_*. Let F_n denote the distribution function defined by $F_n(x) = \hat{F}_n(x; \omega)$ and Q_n the measure defined by $Q_n(A) = \hat{Q}_n(A; \omega)$. Let F be a pointwise limit of $\{F_n\}$. For simplicity in notation we shall assume that $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$. We shall show that

$$\mathcal{L}(F_0) \leq \liminf_{n \to \infty} \mathcal{L}_n(\hat{F}_n)(\omega) \leq \limsup_{n \to \infty} \mathcal{L}_n(\hat{F}_n)(\omega) \leq \mathcal{L}(F).$$

The first inequality follows from $\mathcal{L}_n(\hat{F}_n)(\omega) \geq \mathcal{L}_n(F_0)(\omega)$, a consequence of the definition of the GMLE, and the fact that $\mathcal{L}_n(F_0)(\omega) \to \mathcal{L}(F_0)$ by the choice of ω. Thus we only need to establish the last inequality. For this note that $\mathcal{L}_n(\hat{F}_n)(\omega)$ can be expressed as

$$\int_\Delta \log [F_n(r) - F_n(l)] \, dQ_n(l, r).$$

The desired inequality is thus equivalent to

$$\limsup_{n \to \infty} \int_\Delta \log [F_n(r) - F_n(l)] \, dQ_n(l, r) \leq \int_\Delta \log [F(r) - F(l)] \, dQ(l, r). \qquad (4.1)$$

8

Now fix a positive integer α and a negative integer q. Then

$$\int_\Delta \log\left[F_n(r) - F_n(l)\right] dQ_n(l, r) \le \int_\Delta q \vee \log\left[F_n(r) - F_n(l)\right] dQ_n(l, r)$$

$$\le \sum_{U \in \mathcal{U}_\alpha} M_n(U) Q_n(U),$$

where

$$M_n(U) = \sup_{(l,r) \in \bar{U}} q \vee \log\left[F_n(r) - F_n(l)\right]$$

and $\bar{U}$ is the closure of U. It is easy to check that $M_n(U) = q \vee \log\left[F_n(r_U) - F_n(l_U)\right]$, where $r_U = \sup\{r : (l, r) \in U\}$ and $l_U = \inf\{l : (l, r) \in U\}$. Thus

$$M_n(U) \to M(U) := q \vee \log\left[F(r_U) - F(l_U)\right] = \sup_{(l,r) \in \bar{U}} q \vee \log\left[F(r) - F(l)\right].$$

Also, by the choice of ω, $Q_n(U) \to Q(U)$ for all $U \in \mathcal{U}_\alpha$. Therefore we can conclude that

$$\sum_{U \in \mathcal{U}_\alpha} M_n(U) Q_n(U) \to \sum_{U \in \mathcal{U}_\alpha} M(U) Q(U).$$

Let now

$$m(U) = \inf_{(l,r) \in \bar{U}} q \vee \log\left[F(r) - F(l)\right], \quad U \in \mathcal{U}_\alpha.$$

Using the bound

$$|q \vee \log(x) - q \vee \log(y)| \le e^{-q}|x - y|, \quad 0 \le x, y \le 1,$$

it is easy to verify that

$$M(U) - m(U) \le e^{-q} \sup_{(l,r) \in \bar{U}} \left[F(r_U) - F(r) + F(l) - F(l_U)\right], \quad U \in \mathcal{U}_\alpha.$$

This shows the following.

(1) If $U = \Delta \cap [(q_{i-1}, q_i) \times (q_{j-1}, q_j)]$, then $M(U) - m(U) > 2/\alpha$ implies either $F(q_i) - F(q_{i-1}) > e^q/\alpha$ or $F(q_j) - F(q_{j-1}) > e^q/\alpha$;

(2) if $U = \Delta \cap [[q_i, q_i] \times (q_{j-1}, q_j)]$, then $M(U) - m(U) > 2/\alpha$ implies $F(q_j) - F(q_{j-1}) > e^q/\alpha$;

(3) if $U = \Delta \cap [(q_{i-1}, q_i) \times [q_j, q_j]]$, then $M(U) - m(U) > 2/\alpha$ implies $F(q_i) - F(q_{i-1}) > e^q/\alpha$.

Of course, if U contains only one point, then $M(U) - m(U) = 0$. Using this, we derive

$$\sum_{U \in \mathcal{U}_\alpha} (M(U) - m(U)) Q(U) \le \frac{2}{\alpha} + |q| \sum_{U \in \mathcal{U}_\alpha} Q(U) I[(M(U) - m(U)) > 2/\alpha]$$

$$\le \frac{2}{\alpha} + |q| \sum_{i=1}^{\beta} P(q_{i-1} < L < q_i) I[F(q_i) - F(q_{i-1}) > e^q/\alpha]$$

$$+ |q| \sum_{j=1}^{\beta} P(q_{j-1} < R < q_j) I[F(q_j) - F(q_{j-1}) > e^q/\alpha]$$

$$\le \frac{2}{\alpha} + |q|(1 + \alpha e^{-q}) 2^{1-\alpha}.$$

9

In the last step we use the facts that

$$P(q_{i-1} < L < q_i) + P(q_{i-1} < R < q_i) \le 2\mu((q_{i-1}, q_i)) \le 2^{1-\alpha}$$

and that at most $1 + \alpha e^{-q}$ among the terms $F(q_1) - F(q_0), \ldots, F(q_\beta) - F(q_{\beta-1})$ exceed e^q/α.

Combining the above shows that

$$\limsup_{n \to \infty} \int_\Delta \log\left[F_n(r) - F_n(l)\right] dQ_n(l, r)$$
$$\le \int_\Delta q \vee \log\left[F(r) - F(l)\right] dQ(l, r) + \frac{2}{\alpha} + |q|(1 + \alpha e^{-q})2^{1-\alpha}.$$

The desired inequality (4.1) follows from this by first letting $\alpha \to \infty$ and then $q \to -\infty$.

5. Proof of the Propositions

Fix $\omega \in \Omega_\mu$. Abbreviate $\hat{F}_n(\cdot; \omega)$ by F_n. Let F be a pointwise limit of F_n. Without loss of generality, assume that $\lim_{n \to \infty} F_n(x) = F(x)$ for all x. Set

$$D = \{x \in \mathbb{R} : F(x) \ne F_0(x)\}.$$

Since $\int |F_n - F_0|\, d\mu \to 0$ and μ is a finite measure in view of the assumption $E(K) < \infty$, we have $\mu(D) = 0$.

PROOF OF PROPOSITION 3.1: We need to show that D does not contain regular continuity points of F_0. Let x_0 be a continuity point of F_0. If x_0 belongs to D, then $F(x_0) \ne F_0(x_0)$ and the continuity of F_0 at x_0 and the monotonicity of F and F_0 yield that there exists a positive ϵ such that either $(x_0 - \epsilon, x_0]$ or $[x_0, x_0 + \epsilon)$ is contained in D. Thus either $\mu((x_0 - \epsilon, x_0]) = 0$ or $\mu([x_0, x_0 + \epsilon)] = 0$, and x_0 is not regular. $\square$

PROOF OF PROPOSITION 3.2: Let x_0 be a continuity point of F_0 which is also an interior point of S, the set of support points of μ. Then x_0 does not belong to D; otherwise, there exist, for each $\epsilon > 0$, support points x_1 and x_2 of μ and a positive η such that $(x_1 - \eta, x_1 + \eta)$ is contained in $(x_0 - \epsilon, x_0]$ and $(x_2 - \eta, x_2 + \eta)$ is contained in $[x_0, x_0 + \epsilon)$ and this leads to the contradiction $\mu(D) > 0$. This shows that $F(x) = F_0(x)$ for all continuity points x of F_0 that belong to the interior of S and proves the first part of Proposition 3.2. The second part follows from the first part and the monotonicity of F and F_0. $\square$

PROOF OF PROPOSITION 3.3: Suppose every point of increase of F_0 is strongly regular. We shall show that D does not contain continuity points of F_0. Let x_0 be a continuity point of F_0. If x_0 is a point of increase of F_0, then it is strongly regular and hence regular and cannot belong to D by Proposition 3.1. Suppose now x_0 is not a point of increase of F_0. Then again x_0 cannot belong to D. Otherwise, either $F(x_0) > F_0(x_0)$ or $F(x_0) < F_0(x_0)$ and we shall show that each leads to the contradiction $\mu(D) > 0$. In the first case, $b := \sup\{x : F_0(x) = F_0(x_0)\}$ is a point of increase of F_0,

$b > x_0$ and $F(b-) \geq F(x_0) > F_0(x_0) = F_0(b-)$; thus $[x_0, b) \subset D$ and, since b is strongly regular by our assumption, $\mu(D) \geq \mu((x_0, b)) > 0$. In the second case, $a := \inf\{x \in \mathbb{R} : F_0(x) = F_0(x_0)\}$ is a point of increase of F_0, $a < x_0$ and $F(a) \leq F(x_0) < F_0(x_0) = F_0(a)$; thus $[a, x_0) \subset D$ and, since a is strongly regular by our assumption, $\mu(D) \geq \mu([a, x_0)) > 0$. This shows that D does not contain continuity points of F_0, which is the desired result of Proposition 3.3. $\square$

PROOF OF PROPOSITION 3.6: Make the assumptions of Proposition 3.6. Then D is empty; otherwise, we can use the continuity of F_0 to construct an open interval, that contains a point of increase of F_0 and is contained in D, and arrive at the contradiction $\mu(D) > 0$. Since D is empty, F_n converges to F_0 pointwise and hence uniformly as F_0 is continuous. This proves Proposition 3.6. $\square$

PROOF OF PROPOSITION 3.8: We shall only give the proof in the case $\mu(\{\tau_1\}) > 0$ and $F_0(\tau_2-) = 1$. We shall show that $D \cap [\tau_1, \tau_2] = 0$. This implies that $F_n(x) \to F_0(x)$ for all $x \in [\tau_1, \tau_2]$, and, by the continuity assumption on F_0, this convergence is even uniform on $[\tau_1, \tau_2]$.

It follows from Corollary 2.3 that $F(\tau_1) = F_0(\tau_1)$. This gives the desired result if $F_0(\tau_1) = 1$. Thus assume from now on that $F_0(\tau_1) < 1$. We are left to show that $D_1 = D \cap (\tau_1, \tau_2]$ is empty. If D_1 were not empty, we could use the continuity assumption on F_0, the monotonicity of F_0 and F and $F(\tau_1) = F_0(\tau_1) < F_0(\tau_2-) = 1$ to show that D_1 contains an open interval (a, b) such that $0 < F_0(a) < F_0(b) < 1$ and $\tau_1 < a < b < \tau_2$ and arrive at the contradiction $\mu(D) \geq \mu((a, b)) > 0$.

6. References

Aragon, J. and Eberly, D. (1992). On convergence of convex minorant algorithms for distribution estimation with interval-censored data. *J. of Computational and Graphical Statistics*, 1, 129-140.

Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.*, 26, 641-647.

Becker, N.G. and Melbye, M. (1991). Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV positivity. *Austral. J. Statist.*, 33, 125-133.

Finkelstein, D.M. and Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*, 41, 933-945.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81, 618-623.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. Birkhäuser Verlag, Basel.

Huang, J. (1996). Efficient estimation for proportional hazards models with interval censoring. *Ann. Statist.*, 24, 540-568.

Keiding. N. (1991). Age-specific incidence and prevalence: A statistical perspective (with discussion). *J. Roy. Statist. Soc. Ser. A*, 154, 371-412.

Peto, R. (1973). Experimental survival curve for interval-censored data. *Applied Statistics*, 22, 86-91.

Rudin, W. (1976). Principles of mathematical analysis. McGraw-Hill. New York.

Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *J. Royal Statist. Soc. Ser. B*, 38, 290-295.

van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Ann. Statist.*, 21, 14-44.

Wang, Z. and Gardiner, J. C. (1996). A class of estimators of the survival function from interval-censored data. *Ann. Statist.*, 24, 647-658.

Wellner, J. A. (1995). Interval censoring case 2: alternative hypotheses. In *Analysis of censored data*, Proceedings of the workshop on analysis of censored data, December 28, 1994-January 1, 1995, University of Pune, Pune, India. *IMS Lecture Notes, Monograph Series.* **27**, 271 - 291. H. L. Koul and J. V. Deshpande, editors.

Wellner, J. A. and Zhan, Y. (1997). A hybrid algorithm for computation of the NPMLE from censored data. *JASA*, **92**, 945-959.

Yu, Q, Schick, A., Li, L. and Wong, G.Y.C. (1998a): Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times. To appear in *Canad. J. Statist.*

Yu, Q, Schick, A., Li, L. and Wong, G.Y.C. (1998b): Asymptotic properties of the GMLE with case 2 interval-censored data. *Prob. & Statist. Lett.*, **37**, 223-228.

Anton Schick
Department of Mathematical Sciences
Binghamton University
Binghamton, NY 13902-6000
E-mail: anton@math.binghamton.edu

Qiqing Yu
Department of Mathematical Sciences
Binghamton University
Binghamton, NY 13902-6000
E-mail: qyu@math.binghamton.edu

SELF-CONSISTENT ESTIMATORS OF SURVIVAL FUNCTIONS WITH DOUBLY-CENSORED DATA

By Linxiong Li and Qiqing Yu

Department of Mathematics, University of New Orleans

New Orleans, LA 70148

and

Mathematics Department, State University of New York at Binghamton

Binghamton, NY 13902

Key Words: Asymptotic normality, generalized MLE, strong consistency.

ABSTRACT

The consistency and asymptotic normality of self-consistent estimators (SCE) of survival functions with doubly-censored data have been studied by many authors. However, to the best of our knowledge, expressions of the asymptotic variance of the SCE have not been derived in the literature. In this paper, under the assumption that the survival time and censoring time distributions are discrete with finitely many jump points, an expression and a consistent estimator of the asymptotic variance of the SCE are presented. A proof of the strong consistency of the SCE is also presented. Our simulation studies indicate that the estimate of the asymptotic variance is very close to the true value even with moderate sample sizes and high censoring rates.

1. INTRODUCTION

Doubly-censored data often arise in biometry studies, reliability research and many other fields. We first look at an example.

EXAMPLE 1 Leiderman *et al.* (1973) presented a study on the time needed for an infant to learn to perform a particular task during the first year. The sampled infants were all born within 6 months of the start of the study. At the time of the start of the study, some children had already known how to perform the task; so their observed times were left-censored. Some children learned the task during the time-span of the study, and their ages were recorded. At the end of the study, some of the children had not yet learned the task, and hence their observed times were right-censored.

From the example we see that the time needed to learn to perform the task (called survival time) can be either left-censored (by the time of the start of investigation), observed exactly, or right-censored (by the time of the end of investigation). The times of the start and end of investigation are called censoring times. This censoring scheme is referred to as double censoring and the corresponding observations are said to be doubly-censored.

One of the most important objectives in the study of doubly-censored data is to estimate the underlying distribution function of the random survival time. Turnbull (1974) firstly proposed a self-consistent algorithm for obtaining the SCE of the distribution function with doubly-censored data.

In recent years, Tsai and Crowley (1985) discussed large sample properties of the SCE; Chang and Yang (1987) and Chang (1990) proved the consistency and the asymptotic normality of the SCE, assuming that the underlying distribution is continuous; and Gu and Zhang (1993) established the strong consistency and asymptotic normality of the SCE for arbitrary distributions satisfying certain identifiability condition (see (2.2) below).

However, expressions and estimates of the asymptotic variance of the SCE have not been discussed in the literature. This, we believe, among other things, hinders the theoretical results from practical uses. Some work needs to be done.

To this end we shall in this paper prove the strong consistency of the SCE without the commonly used identifiability condition (Section 3), establish the asymptotic normality and present a consistent estimator of the variance of the SCE (Section 4). A brief numerical investigation of the asymptotic variance is illustrated in Section 5. Next section introduces some necessary definitions and notations.

2. DEFINITIONS AND NOTATIONS

The formal definition of double censorship model is as follows. Let X denote a random survival time with distribution function F, and let (Z,Y) denote a random inspection vector such that $0 \le Z < Y$ with probability one. Denote $G(z,y)$ the distribution function of (Z,Y). Assume that X and (Z,Y) are independent. The observed (doubly-censored) data are $\max\{\min\{X,Y\},Z\}$ and the information that X is either less than Z, between Z and Y, or greater than Y. Thus we can represent such doubly-censored data as a random interval

$$\{L,R\} = \begin{cases} [X,X] & \text{if } Z \le X \le Y \text{ (observed exactly)} \\ (Y,\infty) & \text{if } Y < X \text{ (right censoring)} \\ [0,Z) & \text{if } X < Z \text{ (left censoring)}. \end{cases} \quad (2.1)$$

That is, doubly-censored data can be viewed as a special case of interval-censored data with $L \le R$ and $L < R$ implies either $L = 0$ or $R = \infty$. Denote $Q(l,r)$ the distribution function of the vector (L,R).

Define $K(t) = P\{Z \le X \le Y | X = t\}, t > 0$. The commonly used identifiability condition in the study of double censoring is that

$$K(t) > 0 \text{ for all } t > 0 \quad (2.2)$$

(Chang, 1990; Gu and Zhang, 1993). This condition essentially implies that every value of X is identifiable, which is not always the case in practice. In Example 1, for instance, X is not identifiable beyond the time of the termination of the study. Furthermore, condition (2.2) implies that F is strictly increasing in $[0,\infty)$, i.e., $\mathcal{O}$ contains all identifiable points of X. Under assumption (2.2), $\mathcal{O} = [0,\infty)$.

In this paper, instead of (2.2) we assume that

both X and (Z,Y) are discrete with a finite number of jumps. (2.3)

For (Z,Y), this assumption could be true if the survival time X is examined at discrete times within a limited period of study (Example 1 belongs to this case). A similar assumption has been used by Finkelstein (1986) among others. As for X, if the distribution function F is not discrete, we can always sufficiently approximate F by a discrete distribution with finite support. For instance, in Example 1 we may discretize X, the time to learn how to play the task, into date. Since every individual has limited lifetime, the number of dates (the values X can take on) is finite. Under our assumption the support S_Q of Q has only finitely many points, say

$S_Q = \{((l_i, r_i), 1 \le i \le q\}$. Let $p_i = P\{L = l_i, R = r_i\}$, then $p_i > 0$ and $\sum_{i=1}^q p_i = 1$.

Now we use Turnbull's (1976) self-consistent algorithm to analyze the doubly-censored data presented as (2.1). Of n observed intervals, suppose that there are N_1 $\{l_1, r_1\}$'s, N_2 $\{l_2, r_2\}$'s,..., and N_q $\{l_q, r_q\}$'s with $N_1 + \cdots + N_q = n$. Denote $\{I_i, 1 \le i \le n\}$ the observed intervals. When n is large, for every $1 \le i \le q$, we have $N_i > 0$ with probability one, so WLOG assume $I_i = \{l_i, r_i\}, 1 \le i \le q$. It is seen that I_i is a closed interval if it corresponds to an exact observation and is an open interval if it corresponds to a censored observation, assuming $[0, x)$ is an open interval. Define inner-

most intervals (Peto, 1973), $A_i, i = 1, 2, \ldots, m$ induced by $\{I_i, 1 \le i \le q\}$ such that $A_j \cap I_i$ is either $\emptyset$ or A_j for all i, j.

Let the endpoints of A_i be $\{a_i, b_i\}; i = 1, 2, \ldots, m$, which then satisfy $a_1 \le b_1 \le a_2 \le \cdots \le b_m$. Peto (1973) showed that the generalized maximum likelihood estimator (GMLE) of F only assigns weight to the innermost intervals. Denote $\delta_{ij} = I(A_j \subseteq I_i)$, where $I(B)$ denotes the indicator function of the set B, and denote $\hat{s} = (\hat{s}_1, \ldots, \hat{s}_m)$, the weight assigned by the estimator to the corresponding innermost intervals. By Turnbull's (1976) iterative procedure, the weight $\hat{s}$ is a solution of the following system of self-consistent equations

$$\hat{s}_j = \sum_{i=1}^q \frac{N_i}{n} \frac{\delta_{ij}\hat{s}_j}{\sum_{k=1}^m \delta_{ik}\hat{s}_k}, \quad j = 1, 2, \ldots, m. \qquad (2.4)$$

A solution $\hat{s} = (\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_m)$ of (2.4) is an SCE of $s = (s_1, s_2, \ldots, s_m)$ with $s_i = P\{X \in A_i\}$. Since a GMLE is also an SCE, we only focus on the SCE hereafter. An SCE $\hat{F}(t)$ of $F(t)$ is uniquely defined for $t \in [b_i, a_{i+1})$ by $\hat{F}(b_i) = \hat{F}(a_{i+1}-) = \hat{s}_1 + \cdots + \hat{s}_i$, but is arbitrary for t being in a non-singleton innermost interval. To avoid this ambiguity we define

$$\hat{F}(t) = \begin{cases} \hat{s}_1, & \text{if } 0 = a_1 \le t < b_1 \\ \sum_{j=1}^i \hat{s}_j, & \text{if } t \in [b_i, a_{i+1}) \\ \sum_{j=1}^{i-1} \hat{s}_j + \hat{s}_i \frac{t-a_i}{b_i-a_i}, & \text{if } t \in [a_i, b_i), a_i > 0 \text{ and } b_i \ne \infty \\ 1 - \hat{s}_m, & \text{if } t > a_m \text{ and } a_m < b_m = \infty, \end{cases} \qquad (2.5)$$

where $[a, b)$ is an empty set if $a = b$. This definition of $\hat{F}$ reduces to the product limit estimator if doubly-censored data reduce to right-censored data.

3. THE STRONG CONSISTENCY OF $\hat{F}$

It is seen that if x is such that $P\{X = x\} > 0$ and $K(x) > 0$ then $[x, x] \in S_Q$. In this case, $[x, x]$ constitutes a singleton innermost interval.

Recall that $P\{X \in A_j\} = s_j$, $j = 1, 2, \ldots, m$, and $s = (s_1, s_2, \ldots, s_m)$. We make an additional assumption: $s_i > 0$ for $i = 1, 2, \ldots, m$ throughout the paper.

Now we prove that as $n \to \infty$, $\hat{s}_j \to s_j$ almost surely for $j = 1, \ldots, m$.

Define $\mathcal{D} = \{(d_1, d_2, \ldots, d_{m-1}); d_i > 0, 1 \le i \le m-1, \sum_{i=1}^{m-1} d_i < 1\}$ and

$$L = L(s') = \sum_{i=1}^{q} [p_i \ln(\sum_{j=1}^{m} \delta_{ij} s_j)],$$

where $s' = (s_1, s_2, \ldots, s_{m-1}) \in \mathcal{D}$ and ln denotes the natural logarithm.

Lemma 3.1 $-L(s')$ *is a convex function on the domain* $\mathcal{D}$.

Proof Since $-\ln x$ is convex, $\ln(\sum_{j=1}^{m} \delta_{ij} s_j)$ is also convex for $s' \in \mathcal{D}$. Consequently, $-L(s')$ is a convex function by observing that it is a linear combination of convex functions with positive coefficients. □

It follows from Lemma 3.1 that there exists one and only one point in $\mathcal{D}$ such that L is maximized at that point.

Lemma 3.2 *Under the double censorship model, a point* $s' \in \mathcal{D}$ *maximizes* $L(s')$ *if and only if it is a solution of the system of self-consistent equations*

$$s_j = \sum_{i=1}^{q} p_i \frac{\delta_{ij} s_j}{\sum_{k=1}^{m} \delta_{ik} s_k}, \quad j = 1, \ldots, m. \tag{3.1}$$

The proof of Lemma 3.2 is a modification to the argument of Turnbull (1976) and thus is omitted here.

Li, Watkins and Yu (1997) showed that solving equation (3.1) for s is equivalent to solving the following equation

$$H(x) = \sum_{i=1}^{q} p_i [I(l_i < x < r_i) \frac{H(x) - H(l_i)}{H(r_i) - H(l_i)} + I(r_i \le x)] \tag{3.2}$$

for a distribution function H. In particular, the equivalence between (3.1) and (3.2) means that the measure μ_H induced by the solution H of (3.2)

gives the weight s, the solution of (3.1), to the corresponding innermost intervals. Thus it follows from Lemmas 3.1 and 3.2 that H is uniquely determined by (3.2). To show that $\hat{s} \to s$ it suffices to show the following lemma.

Lemma 3.3 *The distribution F is a solution of (3.2).*

Proof Verify that

$$p_i = P\{L = l_i, R = r_i\} = \begin{cases} P\{Z = r_i\} P\{X < r_i\} & \text{if } 0 = l_i < r_i \\ P\{Y = l_i\} P\{X > l_i\} & \text{if } l_i < r_i = \infty \\ P\{Z \le l_i \le Y\} P\{X = l_i\} & \text{if } l_i = r_i. \end{cases}$$

Thus (3.2) can be written as

$$H(x) = \int_{l < z \le r} \frac{H(x) - H(l)}{H(r-) - H(l)} dQ(l, r) + P\{R \le x\}$$
$$= \int_{0 < z \le x} \frac{H(x) F(z-)}{H(z-)} dG_Z(z)$$
$$+ \int_{y < z < \infty} \frac{[H(x) - H(y)][1 - F(y)]}{1 - H(y)} dG_Y(y) + P\{R \le x\} \tag{3.3}$$

where $G_Y(y)$ and $G_Z(z)$ are marginal distribution functions of Y and Z, respectively. Note that replacing H by F on the right hand side of (3.3) yields

$$\int_{0 < z < x} \frac{F(x) F(z-)}{F(z-)} dG_Z(z)$$
$$+ \int_{y < z < \infty} \frac{[F(x) - F(y)][1 - F(y)]}{1 - F(y)} dG_Y(y)$$
$$= P\{X \le x < Z\} + P\{Y < X \le x\} + P\{R \le x\}$$
$$= P\{X \le x\} \quad (= F(x))$$

since $P\{R \le x\} = P\{X \le x, X \in [Z, Y]\} + P\{X < Z \le x\}$ and

$$P\{X \le x\} = P\{X \le x, X \in [Z, Y]\} + P\{X < Z \le x\}$$
$$+ P\{X \le x \le Z\} + P\{Y < X \le x\}.$$

Thus $H(x) = F(x)$ is a solution of (3.2), which was to be shown. □

The following theorem follows immediately from Lemmas 3.1, 3.2 and 3.3.

Theorem 3.1 *Under the double censorship model, the solution of (3.1) is unique and $\hat{s}$ is strongly consistent. Furthermore, $\hat{F}(x)$ defined by (2.5) is also strongly consistent for $x \in \mathcal{O}$.*

4. THE NORMALITY OF $\hat{F}$

By assumption (2.3), the likelihood function of X can be written as a finite-dimensional parametric model. Thus one may use the observed Fisher information matrix to estimate the variance matrix of the GMLE, provided all regularity conditions in Cramer's theorem are satisfied. Here, however, we are dealing with SCE, and thus a different approach to find the asymptotic variance of the SCE is desired. From Lemmas 3.1 and 3.2 we see that s can be uniquely determined as a function of $p = (p_1, p_2, \ldots, p_q)$, denoted by $s = s(p)$. Let $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_{q-1})$ where $\hat{p}_i = N_i/n$, then $(N_1, \ldots, N_{q-1})$ follows a multinomial distribution $M(n; p_1, \ldots, p_{q-1})$. Denote $\Sigma_{\hat{p}}$ the $(q-1) \times (q-1)$ covariance matrix of $\hat{p}$ where

$$\Sigma_{\hat{p}} = (\alpha_{ij})_{(q-1)\times(q-1)}, \quad \alpha_{ij} = \begin{cases} p_i(1-p_i)/n & \text{if } i = j \\ -p_i p_j/n & \text{otherwise.} \end{cases}$$

Then, it follows from Rao (1973, p.382) that

$$\Sigma_{\hat{p}}^{-1/2}(\hat{p} - p) \xrightarrow{\mathcal{D}} N(O, I), \tag{4.1}$$

where $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution, O is a $(q-1)\times 1$ zero vector, and I is the $(q-1)\times(q-1)$ identity matrix. We now derive the asymptotic distribution of $\hat{s}$ in terms of that of $\hat{p}$.

Since $s_j > 0$ for all j, (3.1) can be written as

$$1 = \sum_{i=1}^{q} p_i \frac{\delta_{ij}}{\sum_{k=1}^{m}\delta_{ik}s_k}, \quad j = 1, \ldots, m. \tag{4.2}$$

Taking partial derivative $\partial s_l/\partial p_h$ on both sides of (4.2) yields

$$0 = \sum_{i=1}^{q} \frac{\partial p_i}{\partial p_h} \frac{\delta_{ij}}{\sum_{k=1}^m \delta_{ik}s_k} - \sum_{i=1}^{q} p_i \frac{\delta_{ij}\sum_{l=1}^m(\delta_{il}\frac{\partial s_l}{\partial p_h})}{(\sum_{k=1}^m \delta_{ik}s_k)^2}$$

$$= \frac{\delta_{hj}}{\sum_{k=1}^m \delta_{hk}s_k} - \frac{\delta_{qj}}{\sum_{k=1}^m \delta_{qk}s_k} - \sum_{i=1}^{q} p_i \frac{\delta_{ij}\sum_{l=1}^{m-1}(\delta_{il}-\delta_{im}/(m-1))\frac{\partial s_l}{\partial p_h}}{(\sum_{k=1}^m \delta_{ik}s_k)^2},$$

or

$$\frac{\delta_{hj}}{\sum_{k=1}^m \delta_{hk}s_k} - \frac{\delta_{qj}}{\sum_{k=1}^m \delta_{qk}s_k}$$
$$= \sum_{l=1}^{m-1}\left\{\left[\sum_{i=1}^{q} p_i \frac{\delta_{ij}(\delta_{il}-\delta_{im}/(m-1))}{(\sum_{k=1}^m \delta_{ik}s_k)^2}\right]\frac{\partial s_l}{\partial p_h}\right\} \tag{4.3}$$

$j = 1, 2, \ldots, m-1$, $h = 1, 2, \ldots, q-1$. Rewrite (4.3) as

$$C\left(\frac{\partial s}{\partial p_1}, \ldots, \frac{\partial s}{\partial p_{q-1}}\right) = (w_1, \ldots, w_{q-1}) \tag{4.4}$$

where

$$C = \left(\sum_{i=1}^{q} p_i \frac{\delta_{ij}(\delta_{il}-\delta_{im}/(m-1))}{(\sum_{k=1}^m \delta_{ik}s_k)^2}\right)_{(m-1)\times(m-1)},$$

$$\frac{\partial s}{\partial p_h} = \begin{pmatrix} \frac{\partial s_1}{\partial p_h} \\ \vdots \\ \frac{\partial s_{m-1}}{\partial p_h} \end{pmatrix}, \quad \text{and} \quad w_h = \left(\frac{\delta_{hj}}{\sum_{k=1}^m \delta_{hk}s_k} - \frac{\delta_{qj}}{\sum_{k=1}^m \delta_{qk}s_k}\right)_{(m-1)\times 1},$$

$h = 1, \ldots, q-1$. Note that the row index in C is j and the column index is l and C is independent of h, while the row index in w_h is j and is a function of h.

By directly taking partial derivatives $\partial^2 s_i/\partial p_j \partial p_h$ on both sides of (4.4), one can see that all the second order partial derivatives exist and are continuous. Therefore, $s = s(p)$ is totally differentiable. Denote

$$\frac{\partial s}{\partial p} = \left(\frac{\partial s_i}{\partial p_j}\right)_{(m-1)\times(q-1)}.$$

By the continuity of the first and second derivatives of s with respect to p, s can be expressed as a function of p using the first order Taylor expansion:

$$\hat{s} - s \approx \frac{\partial s}{\partial p}(\hat{p} - p).$$

Therefore, by the result of multivariate normal convergence theory (see, for example, Rao 1973, p.387), we have the following theorem.

Theorem 4.1 *Let* O *be an* (m-1) × 1 *zero vector and* I *be the* (m-1) × (m-1) *identity matrix. Then*

$$\Sigma_{\hat{s}}^{-1/2}(\hat{s}-s) \xrightarrow{\mathcal{D}} N(O,I) \text{ where } \Sigma_{\hat{s}} = \left(\frac{\partial s}{\partial p}\right)\Sigma_{\hat{p}}\left(\frac{\partial s}{\partial p}\right)^t$$

is the covariance matrix of ŝ.

Let $d_x = \#\{j; a_j \leq x\}$ and e_x be a (m-1) × 1 vector whose first d_x elements are 1 and the rest are 0. Then the consistency of ŝ and Theorem 4.1 imply the next theorem.

Theorem 4.2 *As* n → ∞, *for* x ∈ O, *we have*

$$(e_x^t \Sigma_{\hat{s}} e_x)^{-1/2}(\hat{F}(x) - F(x)) \xrightarrow{\mathcal{D}} N(0,1).$$

By substituting p̂ for p and ŝ for s in $\Sigma_{\hat{s}}$ we obtain an estimator $\hat{\Sigma}_{\hat{s}}$ of $\Sigma_{\hat{s}}$. By the strong consistency of p̂ and ŝ, $\hat{\Sigma}_{\hat{s}}$ is also strongly consistent. Thus we have the following theorem.

Theorem 4.3 *Let* $\hat{\sigma}^2 = e_x^t \hat{\Sigma}_{\hat{s}} e_x$. *Then for* x ∈ O,

$$\frac{\hat{F}(x) - F(x)}{\hat{\sigma}} \xrightarrow{\mathcal{D}} N(0,1).$$

5. NUMERICAL STUDIES

Tables 1 and 2 below are examples of simulation studies we have carried out to assess the closeness between the asymptotic variance given by Theorem 4.3 and the sample variance of F̂. We found that the asymptotic distribution converges quickly to its limiting distribution even with moderate sample sizes. Tables 1 and 2 are identical except different sample sizes. In Table 1 (2), 2000 simulated samples are generated with n=200

Table 1. Standard Deviation of SCE with $n = 200$

x	$F(x)$	$\hat{F}(x)$	$SD(\hat{F}(x))$	$\hat{\sigma}(\hat{F}(x))$
23	0.157			
24 -	0.157	0.160	0.047	0.048
24	0.214	0.213	0.041	0.040
25	0.274	0.272	0.037	0.038
26	0.336	0.335	0.035	0.036
27	0.400	0.400	0.035	0.036
28	0.467	0.466	0.037	0.038
29	0.536	0.535	0.039	0.041
30	0.607	0.606	0.044	0.045
31	0.681	0.679	0.057	0.055

Table 2. Standard Deviation of SCE with $n = 100$

x	$F(x)$	$\hat{F}(x)$	$SD(\hat{F}(x))$	$\hat{\sigma}(\hat{F}(x))$
23	0.157			
24 -	0.157	0.168	0.061	0.067
24	0.214	0.213	0.057	0.057
25	0.274	0.272	0.053	0.053
26	0.336	0.335	0.051	0.051
27	0.400	0.400	0.052	0.052
28	0.467	0.466	0.052	0.053
29	0.536	0.535	0.057	0.057
30	0.607	0.607	0.065	0.064
31	0.681	0.679	0.082	0.077

(100) pairs of (l_i, r_i) in each sample. The observations are obtained by the following scheme:

1. The survival time X follows the distribution

$$F(x) = c_1 \sum_{i \geq 21}^{35} iI(i \leq x), \text{ where } c_1 \text{ is the normalizing constant.}$$

2. The vector (Z, Y) satisfies that the marginal distribution of Z is

$$G_Z(z) = c_2 \sum_{i=24}^{29} i I(i \leq z),$$ where c_2 the normalizing constant,

and $Y = Z + 2$.

Then (L, R) is generated according to (2.1). The censoring rate $P\{X \notin [Z, Y]\} \approx 84\%$, i.e., about 84% of the observations are either right- or left-censored. Here $\mathcal{O} = [24, 31]$ and thus we are not able to correctly estimate $F(t)$ for $t < 24$ or $t > 31$.

The entries in the first column are the all possible values of the survival time X. The second column is the distribution function F of X. The next two columns, $\hat{\bar{F}}(x)$ and $SD(\hat{F}(x))$, give the sample mean and sample standard deviation of the SCE $\hat{F}(x)$ over 2000 repetitions. The last column gives the value computed from the formula given by Theorem 4.3. The entries on the row of $x = 24-$ gives the left-hand limit of corresponding values on that row. We see from columns 4 and 5 that $\hat{\sigma}$ is close to the sample standard deviation of $\hat{F}$.

ACKNOWLEDGEMENTS

BIBLIOGRAPHY

[1] Chang, M.N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. Ann. Statist., 18, 391-404.

[2] Chang, M.N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly censored data. Ann. Statist., 15, 1536-1547.

[3] Finkelstein, D.M. (1986). A proportional hazards model for interval-censored failure time data. Biometrics, 42, 845-854.

[4] Gu, M.G. and Zhang, C-H. (1993). Asymptotic properties of self-consistent estimator based on doubly censored data. Ann. Statist., 21, 611-624.

[5] Leiderman, P.H., Babu, D., Kagia, J., Kraemer, H.C. and Leiderman, G.F. (1973). African infant precocity and some social influences during the first year. Nature, 242, 247-249.

[6] Li, L., Watkins, T. and Yu, Q. (1997). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval-censored data. Scan. J. Statist., to appear.

[7] Peto, R. (1973). Experimental survival curve for interval-censored data. Applied Statist., 22, 86-91.

[8] Rao, C. R. (1973). Linear statistical inference and its applications. New York: Wiley.

[9] Tsai, W. and Crowley, J. (1985). A large sample study of the generalized maximum likelihood estimators from incomplete data via self-consistency. Ann. Statist., 13, 1317-1334.

[10] Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. JASA, 69, 169-173.

[11] Turnbull, B. W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. JRSS B, 38, 290-295.

Dose-Ranging Study of Indole-3-Carbinol for Breast Cancer Prevention

George Y.C. Wong,* Leon Bradlow, Daniel Sepkovic, Stephanie Mehl, Joshua Mailman, and Michael P. Osborne

Strang Cancer Prevention Center, New York, New York

Abstract   Sixty women at increased risk for breast cancer were enrolled in a placebo-controlled, double-blind dose-ranging chemoprevention study of indole-3-carbinol (I3C). Fifty-seven of these women with a mean age of 47 years (range 22–74) completed the study. Each woman took a placebo capsule or an I3C capsule daily for a total of 4 weeks; none of the women experienced any significant toxicity effects. The urinary estrogen metabolite ratio of 2-hydroxyestrone to 16α-hydroxyestrone, as determined by an ELISA assay, served as the surrogate endpoint biomarker (SEB). Perturbation in the levels of SEB from baseline was comparable among women in the control (C) group and the 50, 100, and 200 mg low-dose (LD) group. Similarly, it was comparable among women in the 300 and 400 mg high-dose (HD) group. Regression analysis showed that peak relative change of SEB for women in the HD group was significantly greater than that for women in the C and LD groups by an amount that was inversely related to baseline ratio; the difference at the median baseline ratio was 0.48 with 95 % confidence interval (0.30, 0.67). No other factors, such as age and menopausal status, were found to be significant in the regression analysis. The results in this study suggest that I3C at a minimum effective dose schedule of 300 mg per day is a promising chemopreventive agent for breast cancer prevention. A larger study to validate these results and to identify an optimal effective dose schedule of I3C for long-term breast cancer chemoprevention will be necessary. J. Cell. Biochem. Suppls. 28/29:111–116.   © 1998 Wiley-Liss. Inc.

Key words: chemoprevention; estrogen metabolites; surrogate endpoint biomarker

Indole-3-carbinol (I3C) is a compound present in cruciferous vegetables such as broccoli, Brussels sprouts, cabbage, and cauliflower. This compound has been shown to protect against certain chemical carcinogens, and to induce the enzyme P450A1, which is responsible for the formation of the estrogen metabolite 2-hydroxyestrone [1]. Cell culture experiments have shown that 2-hydroxyestrone acts to block proliferation and inhibit promotion of anchorage independent growth in mouse mammary cells, while its competitive counterpart 16α-hydroxyestrone acts in a promotional manner [2,3]. Therefore, the ratio of 2-hydroxyestrone to 16α-hydroxyestrone, as determined by an ELISA assay [4], is a potential surrogate endpoint biomarker (SEB) for breast cancer prevention. Two animal studies have shown that elevating the estrogen metabolite ratio protects against mammary tumor formation. Bradlow et al. [5] showed this to be the case in the C3HOuJ model, and Grubbs et al. [6] showed this in the DMBA-induced rat model. In the latter case, protection was almost complete. A study in women at various levels of breast cancer risk showed that 16α-hydroxyestrone was elevated in women at greater familial risk for breast cancer [7]. The same phenomenon had been observed in mice at different levels of breast cancer risk [8]. In a recent study, women who had a low metabolite ratio due primarily to the presence of an enzyme defect, which blocks 2-hydroxylation of estradiol, showed a 10-fold increase in breast cancer incidence [9]. The ability of I3C to promote 2-hydroxylation has been demonstrated both in breast cancer cell culture experiments [10,11] and in animal studies [5,6].

The ability of I3C to induce a significant increase in 2-hydroxylation in humans in a short time was first demonstrated by Michnovicz and Bradlow [12]. A 3-month trial of I3C at 400 mg per day against a placebo control and a high fiber diet control showed that the metabo-

lite shift in favor of 2-hydroxylation pathway was sustained over the entire trial period and that no significant adverse effects were observed [13]. The results from these studies suggest that I3C may be a promising chemopreventive agent for breast cancer prevention. We launched a short-term dose-ranging study of I3C in women at increased risk for breast cancer. The overall aim of the intervention study was to determine a minimum effective dose (MED) of I3C, which will not exceed the safely tolerated dose of 400 mg per day established [13] and which will result in a sustained increase in 2-hydroxylation over a 4-week trial period. Five doses of I3C were considered: 50, 100, 200, 300, and 400 mg. A secondary objective of the study was to assess toxicity effects of I3C when taken daily for 4 consecutive weeks. The SEB used in this study was ratio of urinary 2-hydroxyestrone to 16α-hydroxyestrone. Sixty women were recruited in the study, and full compliance was obtained in 57. A placebo-controlled, double-blind trial design was adopted for the study. MED was statistically determined to be 300 mg, and a significant difference was established in the up-regulation of the SEB between the MED group and the placebo group. No significant toxicity effects were observed in the 57 women at the end of the 4-week trial.

STUDY POPULATION

Adult women in good general health but at increased risk for breast cancer were candidates for the dose-ranging study. A woman is considered to be at increased risk for breast cancer either if she is over 60 years of age or she has a family history of the disease (at least one first-degree relative or at least two second-degree relatives with a history of breast cancer). Women who have had a diagnosis of lobular carcinoma in situ or atypia hyperplasia are also considered to be at increased risk in our study.

A number of exclusion criteria were imposed in order to minimize the chances of confounding the outcome of the particular estrogen biomarker chosen. These included thyroid disorders, regular cigarette smoking within the last 6 months, obesity defined as 25% overweight using the nomograph for Body Mass Index, severe anorexia, breast feeding, pregnancy or intention to become pregnant during the study period. In addition, women who have had any form of cancer other than basal or squamous cell carcinoma of the skin, or carcinoma in situ of the cervix, were excluded from the study. Finally, women who regularly consume a large amount of cruciferous vegetables were also excluded because of the nature of our intervention study.

A total of 60 women who were eligible for the trial were selected from over 100 women who were eager to participate in the study. Most of the women came from the New York metropolitan area. Each eligible woman was required to sign an informed consent form before entering the study.

STUDY DESIGN

A placebo-controlled, double-blind design was adopted for the dose-ranging study. Because a rigorous toxicity analysis had not been previously carried out, a dose-escalation scheme was used in the dose assignment for safety considerations. First, ten women in the control group were given placebo capsules. This was followed by assignments of ten women to each of the five ascending dose groups.

A pre-menopausal participant was asked to schedule her appointment within 3 days after her next period ended. Every participant was asked to bring in two first morning urine samples, one from the morning prior to the appointment and the other from the morning of the appointment. A blood sample was taken from each eligible woman on the appointment day and she was given a bottle containing seven capsules of placebo or I3C. One week later, for a total of 4 weeks, a first morning urine sample and a blood sample were collected, and a refill was dispensed for the following week. Because no reliable biochemical tests for I3C metabolites are available, compliance monitoring was carried out by both pill count and an interview.

STATISTICAL METHODS

In the dose-ranging analysis, the level of perturbation of the SEB, namely the urinary estrogen ratio, at any time point was expressed as relative change from baseline. For each dose group, including the placebo group, the peak relative change (PRC) over the 4-week trial period was obtained for each woman, and the mean of the PRC was used to estimate the peak relative perturbation for the particular dose group over the trial period. We remark that a more sensitive approach utilizing a parametric statistical model was not feasible here because

the individual response profiles could not be summarized by a simple parametric curve (for instance, a sigmoidal curve). The estimated PRC from each dose group was then plotted against a dose of I3C to search for an MED. Our dose-ranging study suggests a clear dichotomy of response between a low-dose group involving 50, 100, and 200 mg, and a high-dose group involving 300 and 400 mg; therefore, parametric model fitting at this stage of dose-ranging study to identify an MED was not necessary. Comparisons of PRC among the dose groups were adjusted for confounding factors using linear regression. To ensure no serious statistical biases were introduced into the dose-ranging analysis due to non-randomness of dose assignment, distributions of various factors that could contribute to biases were compared across the three dose groups.

FOOD ITEM ANALYSIS

Every participant was required to complete a simple food intake questionnaire regarding her eating habits in the past 3 months preceding her initial interview for the intervention trial. Both the frequency and the serving size of a variety of vegetables, including most known I3C-rich vegetables, were recorded for each woman. A numeric score representing the total monthly consumption of a specific vegetable item was calculated from the food intake data. Assuming equal weight for every vegetable item, we derived for each woman a I3C vegetable consumption score and the proportion of I3C vegetables in the total vegetables consumed averaging over a month. Data from a total of 54 participants were available for such a food intake analysis. Both the I3C score and the proportion of I3C vegetable consumption were not significantly related to baseline urinary estrogen ratio.

TOXICITY ANALYSIS

Clinical chemistry and complete blood counts were determined from the blood samples collected at baseline and at the end of each of the 4 consecutive weeks of trial. Any parameter whose measured value was outside the normal range was investigated for possible toxicity. Except for two participants who had unexplained small increases in the liver enzyme SGPT level (43 to 65, and 30 to 71), no other toxicity effects were encountered.

DOSE-RANGING ANALYSIS

A total of 57 women were evaluable for the entire dose-ranging study. Except for three women from New Jersey, all of the 57 women were from the New York metropolitan area. Fifty-two (91%) of the women were white. Forty-six (81%) were college educated, and twenty-four (42%) completed graduate studies. The average age of the participants was 46.7 years (range 22–74). The average age at menarche was 12.4 years (range 8–18). Forty (70%) of the women were pre-menopausal, and 38 (67%) of the women have been pregnant at least once.

Figure 1 displays the sample mean relative change of the estrogen ratio from baseline over time for the control group (n = 10), 50 mg group (n = 7), 100 mg group (n = 10), 200 mg group (n = 10), 300 mg group (n = 10), and 400 mg group (n = 10). The profiles suggest a segregation of the treated groups into a low-dose group (LD) consisting of women in the 50, 100, and 200 mg groups, and a high-dose (HD) group consisting of women in the 300 and 400 mg groups. Moreover, the plots also suggest that the control group (C) was not significantly different from the LD group. For the sake of statistical power, the dose-ranging analysis hereafter will compare data from the C, LD, and HD groups.

Before we can statistically compare the levels of perturbation of the SEB in the three groups, we have to rule out the presence of any statistical bias due to non-randomness of dose assignments to the participants. To this end, we examined the distributions of a number of potential
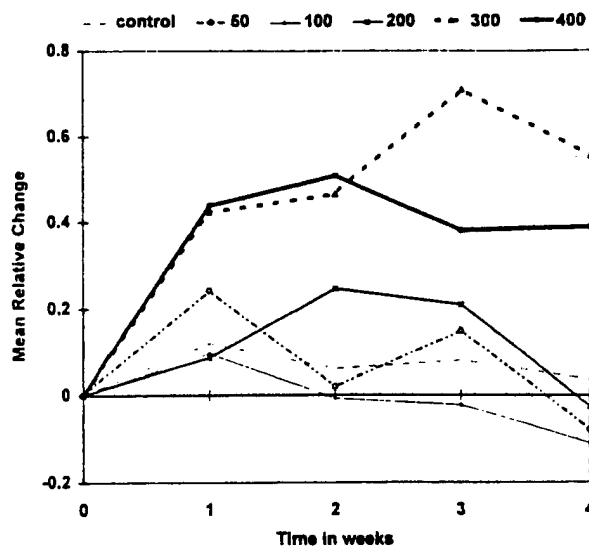

Fig. 1. Mean relative change of urinary estrogen ratio profile plots for the control and five dose groups.

confounding factors, including age, age at menarche, baseline estrogen ratio, menopausal status, pregnancy history, and educational level. No significant differences were found across the three groups with respect to such factors.

Within each of the three groups, we identified the PRC for each of the participants in the group and calculated the usual 95% confidence interval (CI) for the population mean PRC for the group. Figure 2 presents the individual PRC values and the CI for each group. There was no significant difference in mean PRC between C and LD. The sample mean $\pm$ SD of PRC for LD was 0.33 ± 0.36 and that for HD was 0.81 ± 0.57; the difference of 0.48 was significant at $P = 0.001$ by the two-sample t-test. The 95% CI for the difference in mean PRC between the HD and LD groups was estimated to be (0.22, 0.76).

The perturbation results were unadjusted for any confounding factors. Menopausal status was a major concern in the comparison. Figure 3a,b shows that within each of HD group and C + LD group, there was no significant difference in mean relative change of the SEB from baseline between pre-menopausal and post-menopausal women over the entire trial period. The same conclusion was true for comparison based on PRC. Besides menopausal status, we also included age, age at menarche, baseline estrogen ratio, and educational level in a multivariate

Fig. 3. a,b: Mean relative change of urinary estrogen ratio profile plots stratified by menopausal status. Vertical bars represent usual 95% confidence intervals for the mean. No significant difference in mean relative change between pre-menopausal and post-menopausal women was established within both control + low-dose group and high-dose group.

regression to attempt to explain the variation in the observed PRC. For the C + LD group, the variation in PRC could only be explained by random inter-participant differences. However, for the HD group, about 50% of the total variation in PRC was significantly explained by a regression towards the mean effect of baseline estrogen ratio ($P = 0.001$). Figure 4 displays the linear relationship between PRC and baseline ratio for the HD group, and the lack of correlation in the case of the C + LD group.

From regression analysis, we found a significant adjusted difference in PRC between the two groups as long as baseline estrogen ratio was less than 2.92. Table I tabulates the differences and the corresponding 95% CIs for some selected values of baseline ratio.
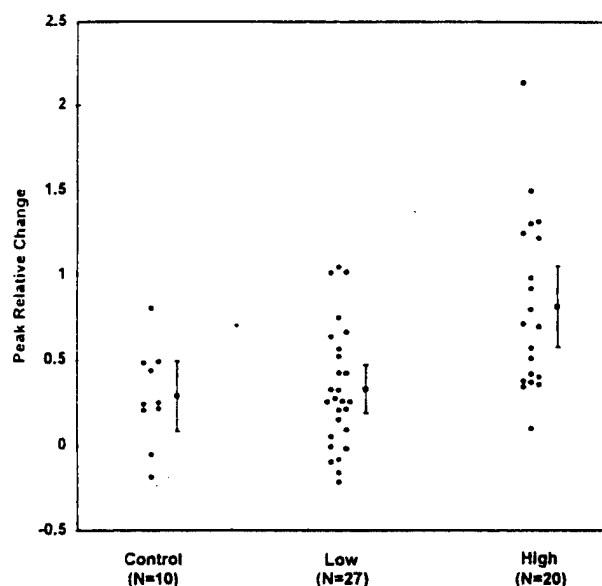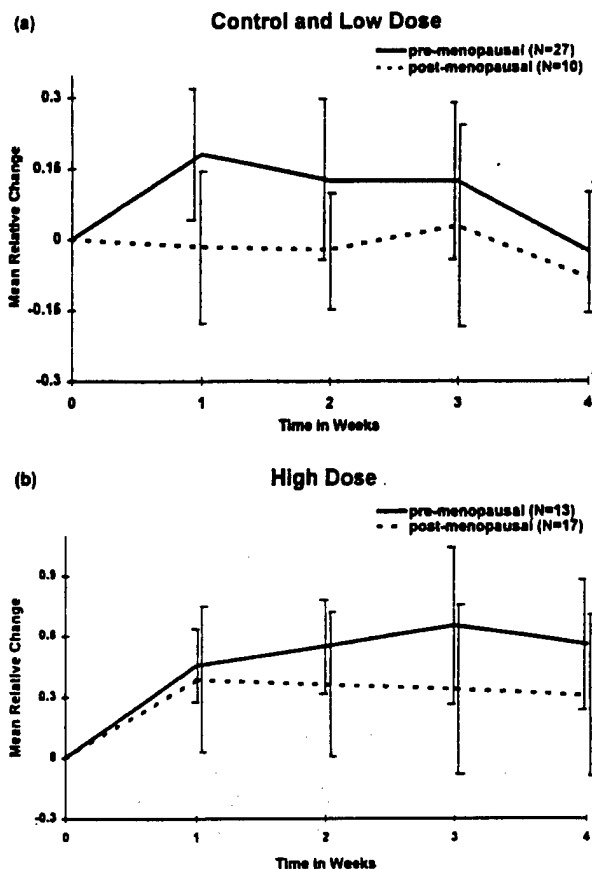
Fig. 2. Comparison of peak relative change of urinary estrogen ratio among control, low- and high-dose groups. Difference between the high-dose group and the other two dose groups, unadjusted for confounding factors, was significant at $P = 0.05$.

DISCUSSION

The goal of this placebo-controlled, double-blind study was to determine a minimum effec-

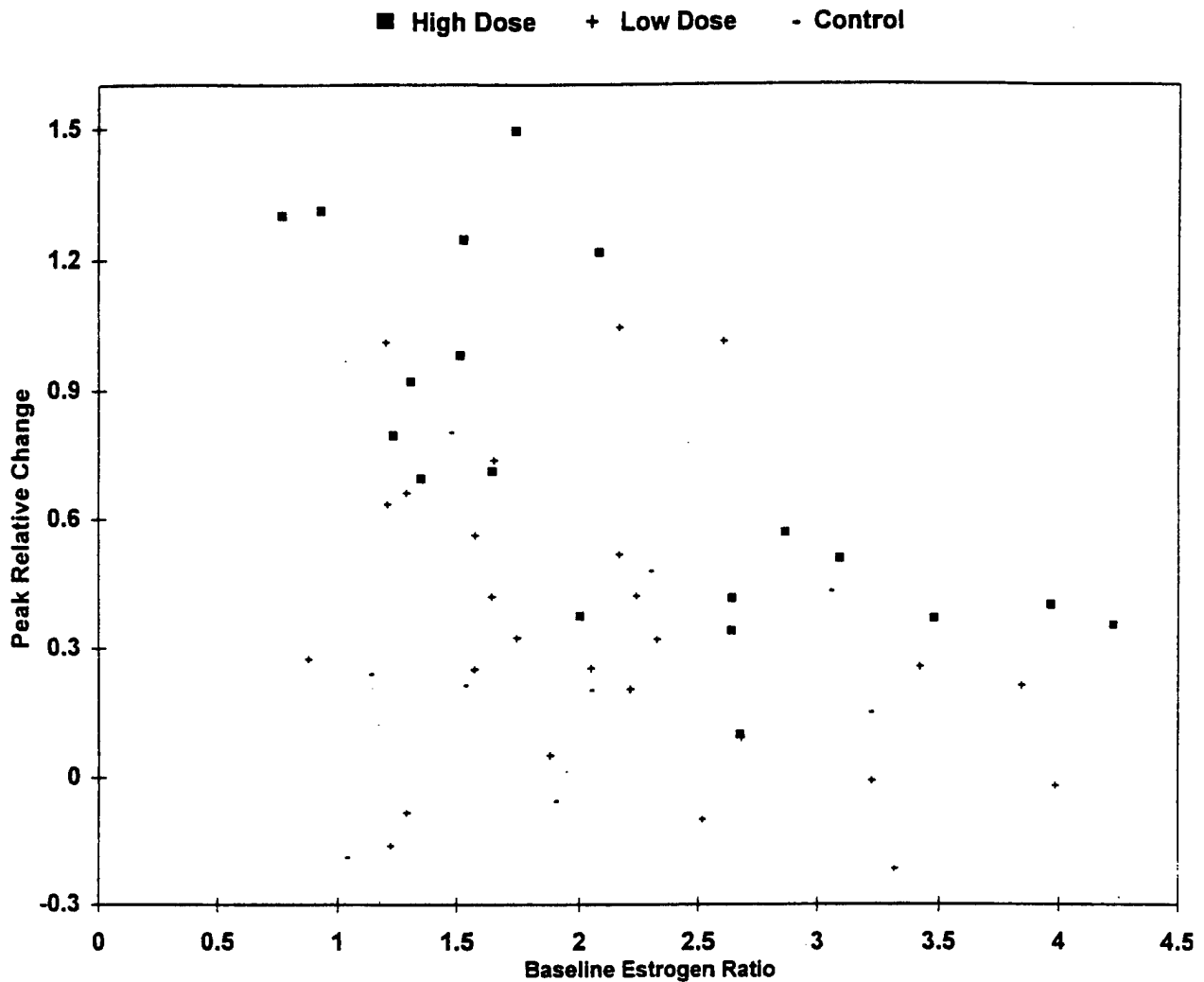■ **High Dose**   + **Low Dose**   - **Control**

Fig. 4. Plot of peak relative change of urinary estrogen ratio vs. baseline value. Linear regression was significant only in the high-dose group: $y = 1.38-0.29x$, $P = 0.001$, $R^2 = 0.50$. For control + low-dose group, mean peak relative change $\pm SD = 0.32 \pm 0.33$.

tive and safe dose schedule of I3C that will result in a significant increase in the urinary estrogen metabolite ratio of 2-hydroxyestrone to 16α-hydroxyestrone. We have shown in a sample of 57 women that an appropriate choice of MED was 300 mg and that daily intake of I3C at this dose presented no significant toxicity in a 4-week trial. At this MED dose schedule, peak relative change of the estrogen metabolite ratio was significantly greater than that at the lower doses, and the difference was more pronounced for women with lower baseline ratios. However, there was no significant perturbation of the biomarker for women with high and presumably already protective baseline ratios. Menopausal status was not a significant factor for perturbation of the biomarker in our analysis, although there was a trend towards greater up-regulation of the ratio in the

TABLE I. Adjusted Differences in PRC of Urinary Estrogen Ratio Between High-Dose Group and Combined Control and Low-Dose Group*

Baseline ratio	Adjusted difference	95% CI Lower	95% CI Upper	P value
Q1 = 1.41	0.65	0.44	0.88	<0.001
M = 2.01	0.48	0.3	0.67	<0.001
Q3 = 2.66	0.29	0.1	0.49	0.004
C = 2.92	0.22	0˙	0.43	0.05

*Q1, M, and Q3 represent the first quartile, median and third quartile of baseline ratio, respectively. C represents the critical baseline ratio beyond which there was so significant difference in PRC between the two groups.

case of pre-menopausal women. A larger study should be conducted to confirm the findings reported here, particularly the lack of effect of menopausal status on the perturbation of the

biomarker, and to identify an optimal effective dose schedule for a long-term breast cancer prevention trial.

REFERENCES

1. Wattenberg LW, Loub WD (1978): Inhibition of polycyclic aromatic hydrocarbon induced neoplasia by naturally occurring indoles. Cancer Res 38:1410–1413.
2. Telang NT, Suto A, Wong GYC, Osborne MP, Bradlow HL (1992): Induction by estrogen metabolite 16α-hydroxyestrone of genotoxic damage and aberrant proliferation in mouse mammary epithelial cells. JNCI 84:634–636.
3. Suto A, Bradlow HL, Wong GYC, Osborne MP, Telang NT (1993): Experimental down-regulation of intermediate biomarkers of carcinogenesis in mouse mammary epithelial cells. Breast Cancer Res Treat 27:193–202.
4. Bradlow HL, Sepkovic DW, Klug T, Osborne MP (1998): Application of an improved Elisa assay to the analysis of urinary estrogen metabolites. Steroids (in press).
5. Bradlow HL, Michnovicz JJ, Telang NT, Osborne MP (1991): Effects of dietary indole-3-carbinol on estradiol metabolism and spontaneous mammary tumors in mice. Carcinogenesis 12:1571–1574.
6. Grubbs C, Steele VE, Casebolt T, Juliane MM, Eto I, Whitaker LM, Dragnec KH, Kelloff GJ, Lubet RL (1995): Chemoprevention of chemically induced mammary carcinogenesis by indole-3-carbinol. Anti-Cancer Res 15:709–716.
7. Osborne MP, Karmali RA, Hershcopf RJ, Bradlow HL, Kourides IA, Williams WR, Rosen PP, Fishman J (1988): Omega-3 fatty acids: Modulation of estrogen metabolism and potential for breast cancer prevention. Cancer Invest 6:629–631.
8. Bradlow HL, Hershcopf RJ, Martucci CP, Fishman J (1985): Estradiol 16α-hydroxylation in the mouse correlates with mammary tumor incidence and presence of murine mammary tumor virus: A possible model for the hormonal etiology of breast cancer in humans. Proc Natl Acad Sci USA 82:6295–6299.
9. Taioli E, Bradlow HL, Garbers SV, Sepkovic DW, Osborne MP, Trachman J, Ganguly S, Garte SJ (1998): Cyp1A1 genotype, estradiol metabolism and breast cancer in African-Americans. Cancer Prev Detect (in press).
10. Niwa T, Swaneck G, Bradlow HL (1994): Alterations in estradiol metabolism in MCF-7 cells induced by treatment with indole-3-carbinol and related compounds. Steroids 58:523–527.
11. Tiwari RK, Li Guo, Bradlow HL, Telang NT, Osborne MP (1994): Selective responsiveness of human breast cancer cells to indole-3-carbinol, a chemopreventive agent. JNCI 86:126–131.
12. Michnovicz JJ, Bradlow HL (1990): Induction of estradiol metabolism by dietary indole-3-carbinol in humans. JNCI 82:947–949.
13. Bradlow HL, Michnovicz JJ, Halper M, Miller DG, Wong GYC, Osborne MP (1994): Long-term responses of women to indole-3-carbinol or a high fiber diet. Cancer Epidemiol Biomarkers Prev 3:591–595.
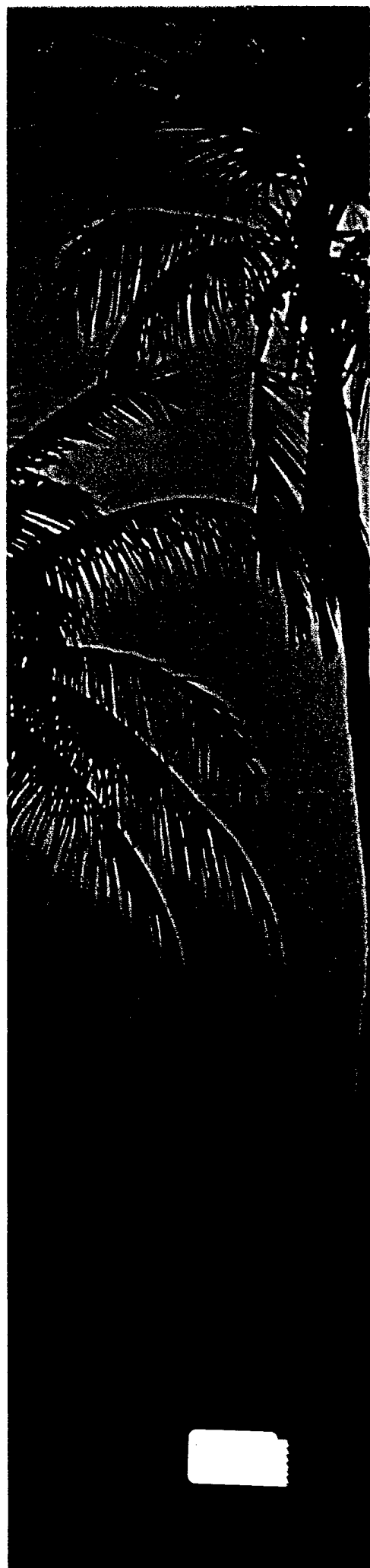
1995 ABSTRACTS

Summaries of Papers
Presented at the
Joint Statistical Meetings

AMERICAN STATISTICAL ASSOCIATION
155th Annual Meeting

INTERNATIONAL BIOMETRIC SOCIETY
Eastern North American Region
Western North American Region

INSTITUTE OF MATHEMATICAL STATISTICS

Orlando, Florida
August 13-17, 1995

ESTIMATION OF A SURVIVAL FUNCTION WITH INTERVAL-CENSORED DATA, A SIMULATION STUDY ON THE REDISTRIBUTION-TO-THE-INSIDE ESTIMATOR

Qiqing Yu, George Wong, Ling Ye

Qiqing Yu, Department of Applied Math and Statistics, SUNY at Stony Brook, Stony Brook, NY 11794

We consider nonparametric estimation of a survival function with interval-censored data. Yu and Wong Propose (1994) a Redistribution-to-the-Inside Estimator (RTIE) to estimate the survival function. The RTIE has an explicit expression and has been shown to be the GMLE and an consistent estimate in some special cases. In this note, we present the results of simulation study on the RTIE.

43. EMPIRICAL BAYES METHODS FOR COMBINING LIKELIHOODS
(Abstracts not available at press time.)

44. APPLICATIONS OF GENERALIZED LINEAR MODELS

VARIABLE SELECTION IN AUTO-LOGISTIC MODELS

Fred W. Huffer, Hulin Wu

Fred W. Huffer, Dept. of Statistics, Florida State University, Tallahassee, FL 32306

KEY WORDS: Spatial Binary Data, Pseudo-likelihood Estimation, Logistic Regression

Besag's auto-logistic model is widely applicable to the modeling of spatial binary data with covariates. However, fitting this model by maximum likelihood (via Markov Chain Monte Carlo) is very time consuming. For this reason maximum likelihood is not practical in the preliminary model-building/covariate selection stage of data analysis if there are many covariates under consideration. Maximum pseudo-likelihood (MPL) estimation is not efficient, but is rapidly computed and easily implemented by standard logistic regression software. We study the use of MPL for covariate selection. In particular, we examine the properties of the "reduction in deviance" and AIC (Akaike's Information Criterion) when these are computed from the pseudo-likelihood function. These quantities often behave somewhat differently than their likelihood-based counterparts and require appropriate adjustment. We apply our results to the selection of good climate covariates for use in modeling the distribution of various Florida plant species.

ESTIMATION OF CANCER MORTALITY RATES : A HIERARCHICAL BAYES GLM APPROACH

Malay Ghosh, Kannan Natarajan, Lance Waller

Malay Ghosh, University of Florida, 223 Griffin-Floyd Hall, Gainesville, Florida 32611

The paper considers estimation of cancer mortality rates for local areas. The raw estimates are usually based o small sample sizes, and hence are usually unreliable. A hierarchical Bayes generalized linear models approach is take which connects the local areas, thereby enabling one to " borrow strength ". Two sets of data are analyzed. The first s of data relates to cancer mortality estimation for several small counties in Missouri where no spatial effects are presen The second set of data relates to cancer mortality estimation for several census tracts in a certain region in the state New York, where spatial effect is present.

SMALL AREA INFERENCE FOR BINARY VARIABLES USING HIERARCHICAL LINEAR MODELS

Donald Malec, J. Sedransk

Donald Malec, ORM, National Center for Health Statistics, 6525 Belcrest Rd., Hyattsville, MD 20782

KEY WORDS: Bayesian Methodology, Disability, Gibbs Sampler, Synthetic Estimation

Using conventional methods, the National Health Interview Survey (NHIS) is designed to provide prec estimates for the entire United States but not for most states nor for subpopulations within states. Our investigati concerns improved inference for binary random variables for small areas and subpopulations within small areas. In t paper, model-based state estimates of the proportion of individuals experiencing partial work limitation (i.e., disabili are developed using Bayesian methods with hierarchical structure. To evaluate the accuracy of these estimates, disabil data collected in the "long form" of the 1990 U.S. Census of Population and Housing are used. First, a sample of Census is drawn to mimic the multi-stage design of the NHIS. Then, data from this sample are used to provide estima for a set of small areas. Finally, these estimates are compared with synthetic estimates and the true values obtained fi the Census.

45. ANALYSIS OF VARIANCE
(Abstracts not available at press time.)

1996 Program

American Statistical Association
156th Annual Meeting

Institute of Mathematical Statistics

International Biometric Society
Eastern North American Region
Western North American Region

1996 JSM
CHICAGO

from x, $g()$ is a known link function and $\beta()$ is a function in u. Conventionally, u are some metrically scaled variables while x are qualitative regressors.

Parameter estimation in varying coefficient models can be done by a local likelihood approach (Tibshirani & Hastie [*J.A.S.A.* 82 (1987):559–568] which is directly feasible by standard software for generalized linear models. Theoretical results yield asymptotic consistency of the estimates and allow for asymptotic pointwise confidence bands. Moreover, a direct correction of the estimation bias is available by using a simple Fisher scoring routine.

The theoretical results are supported by simulations and real data examples.)

126.   SURVIVAL ANALYSIS I

ON SEMIPARAMETRIC RANDOM CENSORSHIP MODELS

Gerhard Dikta

Fachhochschule Aachen, Abteilung J\"ulich, Ginsterweg 1, D - 52428 J\"ulich, Germany DIKTA@FHSERVER03.DVZ.FH-AACHEN.DE

In the random censorship model one observes data of the form (Z,δ) where $Z=(X, Y)$, X is independent of Y, and δ indicates whether X is censored ($\delta=0$) or not ($\delta=1$). Denote by $m(x)=IE(\delta\backslash Z = x)$ the regression function of the binary datum δ given $Z = x$ and assume that m belongs to a parametric family with parameter space $\Theta \subset IR^k$, i.e. $m(x) = m(x, \theta_0)$ and $\theta_0 \in \Theta$. We propose a semiparametric estimator of the distribution function F of X, denoted by F_n, which is based upon maximum likelihood estimation of θ_0, and which generalizes the Cheng and Lin estimator in the proportional hazards model. We establish uniform consistency and a functional central limit result for F_n which is compared to that of the Kaplan-Meier estimator.

VARIANCE OF THE MLE OF A SURVIVAL FUNCTION WITH DOUBLY-CENSORED DATA

Qiqing Yu, Linxiong Li and George Wong

Qiqing , SUNY at Binghamton, University of New Orleans, and Strang Cancer Preventive Institute

The asymptotic properties of the nonparametric MLE or the self-consistent estimator of a survival function with doubly-censored data have been studied by many authors. However, to date, it is not clear from the literature how to produce an estimate of the asymptotic variance of the MLE of $S(t)$ with doubly-censored data, even though the existence of such asymptotic variance has been proved, with an abstract form in the Banach space (Gu and Zhang (1993)). We present the explicit expressions of the asymptotic variance of the generalized MLE and its estimator.

Simulation study indicates that the approximation is close even with sample size $n=100$ and the probability of censoring is 85%.

DOUBLE CENSORING: CHARACTERIZATION AND COMPUTATION OF THE NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATOR

Jon A. Wellner and Yihui Zhan

Yihui Zhan, University of Washington, Department of Statistics, Box 354322, Seattle, WA 98195. Email: zhan@stat.washington.edu

KEY WORDS: Double Censoring, NPMLE, ICM Algorithm, Hybrid Algorithm

While the likelihood equations have a unique solution in the case of right censored data, this is no longer the case for doubly censored data: the likelihood equations may have multiple solutions in the case of double censoring. Algorithms such as the EM algorithm designed to calculate one solution of the likelihood equations may converge to a self-consistent estimate other than the NPMLE. The ambiguity of the EM algorithm in calculating the NPMLE for doubly censored data and its known slow convergence rate pose real difficulties in applications, especially when bootstrap methods are used for inference.

In this paper we present a characterization of the NPMLE for doubly censored data. The NPMLE is characterized as the left-derivative of a convex minorant formed by derivatives of likelihood function. The NPMLE is shown to be one of the self-consistent estimates maximizing the likelihood function. Based on the characterization, we propose a new hybrid algorithm that utilizes a composite algorithmic mapping of the EM algorithm and the modified ICM algorithm. Numerical simulations demonstrate that the hybrid algorithm converges to the NPMLE more rapidly than either of the EM or the naive ICM algorithm for doubly censored data.

PROPERTIES OF TEST STATISTICS APPLIED TO RESIDUALS IN FAILURE TIME MODELS

Inmaculada B. Aban, Edsel A. Peña

Inmaculada B. Aban, Department of Math (084), University of Nevada Reno, Reno, NV 89557

KEY WORDS: Generalized Residual Process, Goodness-of-Fit, Model Validation

Asymptotic properties of a class of test statistics when applied to hazard-based residuals arising in survival and reliability models will be presented. These test statistics are useful in goodness-of-fit testing and model validation. The properties are obtained by examining the asymptotic properties of generalized residual processes, which are (possibly random) time-transformations of the processes associated with the incomplete failure times. Since the time-transformations depend on unknown model parameters, the residual processes are obtained by replacing the unknown parameters by their estimators. The results therefore shed light on the effects of estimating parameters to obtain the residual processes. Implications concerning possible pitfalls of some existing model validation procedures utilizing hazard-based residuals and ways to correct these problems will be discussed.

SISC-96

Sydney International Statistical Congress

Sydney - Australia
July 8 - 12, 1996
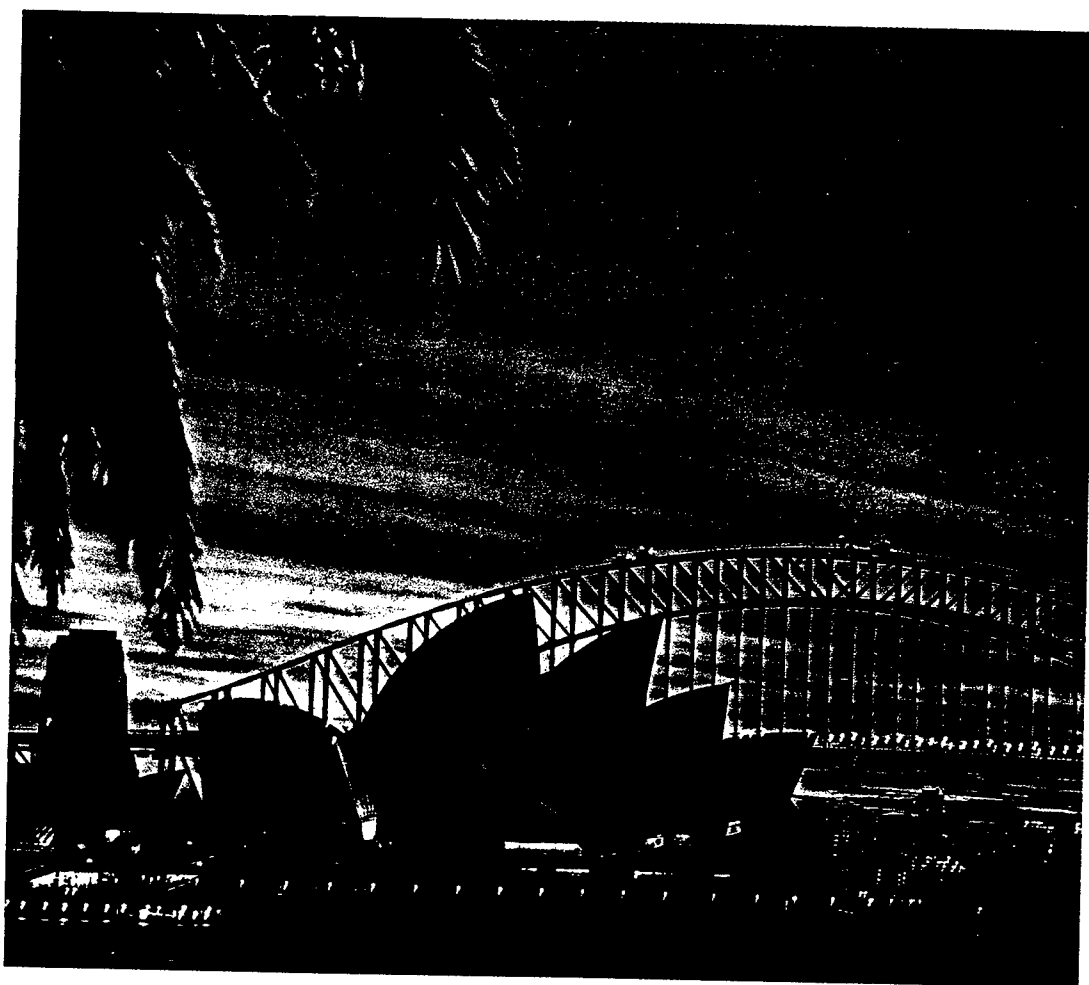
13th Australian Statistical Conference
   Computing Science and Statistics: 28th Symposium on the Interface
IMS Special Topics Meeting

Interface '96

Final Program

ASC/IMS Contributed: Topics in Statistical Inference III

320 The Behaviour of the Maximum Likelihood Estimator as a Process and Some Applications

Robert M. LOYNES
University of Sheffield, England.

Given a set of observations, supposedly either independent and identically distributed or from a stationary AR process, whose distribution contains a fixed-dimension unknown parameter, the behaviour of the maximum–likelihood estimator (MLE) as a function of the number of observations used contains evidence about whether the model assumptions are satisfied, or whether a change of regime or drift is taking place. A weak convergence result for the process of MLEs is given, which allows various tests to be constructed. [✍ Prof R.M. Loynes, University of Sheffield, Probability & Statistics Section, School of Maths & Stats, Sheffield S3 7RH UK; R.LOYNES@SHEFFIELD.AC.UK.]

321 Comparing Groups with Irregular Longitudinal Data

J. S. MARITZ
Medical Research Council, South Africa.

Longitudinal data arise when observations of a dependent variable are made at several successive time points. When such data are recorded for a number of subjects it often happens that the time configuration varies from subject to subject, producing irregular longitudinal data. Comparison of two or more groups of subjects is considered using exact permutational methods. This entails choosing appropriate descriptive and test statistics and generating their exact distributions. [✍ J. S. Maritz, MRC-CERSA, PO Bo 19070, Tygerberg 7505, South Africa; SMARITZ@EAGLE.MRC.AC.ZA.]

322 Repeated Ordinal Responses

Rory St John WOLFE
Southampton University, UK.

An approach to modelling repeated ordinal responses is discussed. This involves using 'scaling' terms in a cumulative logit model [McCullagh *J. Roy. Statist. Soc.Ser.B* 42 1980:109–142]. The approach is applied to data from telecommunication experiments. A new general purpose method of fitting the model in GLIM4 is introduced. Finally the consideration of a random-effects model is discussed. [✍ Rory Wolfe, Maths Department, Southampton University, Highfield, Southampton, SO17 1BJ, UK; RW@MATHS.SOTON.AC.UK.]

323 On Minimum Distance Estimation of Location Parameter for Interval Censored Data

Vasudaven MANGALAM
Curtin University, Perth, Australia.

Let $X_1, X_2, ... X_n$ be i.i.d. with distribution function given by $F(x - a)$ where F is an unknown symmetric distribution and a is an unknown location parameter. $T_1, T_2, ... T_n$ are i.i.d. and independent of $X_i's$ with unknown distribution G. We observe $(T_i, d_i), i = 1, ..., n$ where d_i is the indicator of whether X_i is less than or equal to T_i. A minimum distance estimator is constructed for the parameter and the properties are studied. Two-sample extension to this is also considered. [✍ Vasudaven Mangalam, School of Mathematics and Statistics, Curtin University of Technology, GPO Box U1987, Perth WA 6001; VASU@CS.CURTIN.EDU.AU.]

324 Variance of the MLE of a Survival Function with Interval Censored Data

Qiqing YU
SUNY at Binghamton.
Linxiong LI
University of New Orleans.

George Y. WONG
Strang Cancer Preventive Institute.

Interval-censored data consist of n pairs of observations (l_i, r_i), $i = 1, ..., n$, where $l_i \leq r_i$. We either observe the exact survival time X if $l_i = r_i$ or only know $X \in (l_i, r_i)$ otherwise. We established the asymptotic normality of the nonparametric MLE of a survival function $S(t)$ ($= P(X > t)$) with such interval-censored data and present an estimate of the asymptotic variance of the MLE. We show that the convergence rate in distribution is in $\sqrt{n}$. Simulation study also supports our result. An application to the cancer research is presented. [✍ Qiqing Yu, Math Department, SUNY at Binghamton, NY 13902, USA; QYU@MATH.BINGHAMTON.EDU.]

325 On the Relationship Between Power of a Test and Shape of its Critical Region

Daryl W. TINGLEY
Maureen A. TINGLEY
University of New Brunswick, Fredericton, NB, Canada.

With the Neyman-Pearson Lemma as yard-stick for comparisons, the relationship is investigated between test power and shape of critical region, as test size approaches zero. Measure-theoretic definitions are used to quantify the notion of similar versus dissimilar critical regions. Results are obtained for two extremes of limiting power: power approaching that of Neyman-Pearson, and power negligible when compared with Neyman-Pearson, as test size decreases. Examples illustrate that small test sizes are not practical when sample estimates replace values of nuisance parameters. [✍ Maureen Tingley, Dept of Math and Stat, University of New Brunswick, Box 4400 Fredericton, NB, Canada E3B 5A3; MAUREEN@MATH.UNB.CA.]

326 A Generalisation of Cochran's Theorem and Its Applications in the Analysis of Variance of Repeated Measures

Júlia T. FUKUSHIMA
San Paulo University.
Regina C. C. P. MORAN
State University of Campinas.
Ioannis G. VLACHONIKOLIS
Loughborough University of Technology, UK.

Cochran's theorem and its many corollaries and interrelationships have played a most prominent role in statistics. The present work attempts to formulate a unified approach by means of two theorems on necessary and sufficient conditions under which the sums of squares of the various hierarchical layers in ANOVA are distributed like multiples of chi-square variables. Some new results concerning the standard univariate F-tests in the analysis of repeated measurements are derived as a special case. [✍ Ioannis G. Vlachonikolis, University of Loughborough, Department of Mathematical Sciences, Loughborough, LEICS LE113TU, UK; I.G.VLACHONIKOLIS@LUT.AC.UK.]

Thursday 11 July: 10:30–12:20

ASC Invited: Session in Celebration of Ted Hannan's Contributions to Time Series – II

327 Estimation of Speed, Direction and Structure from Spatial Array Data

David R. BRILLINGER
University of California, Berkeley, USA.

The contributions of E.J. Hannan and his collaborators to the problem of the estimation of speed, direction and structure from spatial array data will be reviewed. One has in mind data such as earthquake signals recorded as the energy passes across an array of seismometers. From such data one can estimate parameters of the earthquake and also of the medium through which

1997 Program of the Joint Statistical Meetings

American Statistical Association
157th Annual Meeting

Institute of Mathematical Statistics

International Biometric Society
Eastern and Western North American Regions

Statistical Society of Canada

Anaheim, California
August 10—14, 1997

This Program Book is included in the registration fee for the 1997 Joint Statistical Meetings.

Cover Photos: James Blank and Robert Shangle, *Beautiful Orange County*

228 Palos Verdes B
ESTIMATION AND ASYMPTOTICS—Regular Contributed Papers
IMS
Chair: Thomas E. Nichols, Carnegie Mellon U
(8:35) ASYMPTOTIC PROPERTIES OF SELF-CONSISTENT
ESTIMATORS WITH DOUBLY-CENSORED DATA. Qiqing Yu, State U of
New York-Binghamton; Linxiong Li, U of New Orleans
(8:50) ON CONSISTENCY OF THE BEST-R-POINTS-AVERAGE
ESTIMATOR FOR THE MAXIMIZER OF A NONPARAMETRIC REGRES-
SION FUNCTION. Z.D. Bai, Mong-Na L. Huang, Nat'l Sun Yat-Sen U
(9:05) A GENERAL ESTIMATION METHOD USING SPACINGS.
Kaushik Ghosh, S. Rao Jammalamadaka, U of California-Santa Barbara
(9:20) ASYMPTOTICS FOR MULTIVARIATE T STATISTIC. Steven J.
Sepanski, Saginaw Valley State U
(9:35) TWO CENTRAL LIMIT THEOREMS FOR FUNCTIONAL
Z-ESTIMATORS. Yihui Zhan, MathSoft, Inc
(9:50) DENSITY ESTIMATION FOR A CLASS OF STATIONARY
NONLINEAR PROCESSES. Kamal C. Chanda, Texas Tech U
(10:05) FLOOR DISCUSSION

College Bowl—10:30 a.m. - 12:20 p.m.

229 H-California B
COLLEGE BOWL SEMIFINALS AND FINALS
Mu Sigma Rho
Organizers: Don Edwards, U of South Carolina; Mark E. Payton,
Oklahoma State U
Chair: Mark E. Payton, Oklahoma State U
Emcee: George Casella, Cornell U
Scorekeeper: Bruce Collings, Brigham Young U
Teams: Winners of Quarterfinals (Session 146)

Invited Sessions—10:30 a.m. - 12:20 p.m.

230 H-California C
CLASSIFICATION OF RACE AND ETHNICITY: A DISCUSSION—
Invited Panel
*Council of Professional Assn on Fed Stat, Sec. on Epidem., Govt. Stat. Sec., Sec.
on Hlth. Policy Stats., Social Stat. Sec.*
Chair/Organizer: Edward J. Spar, COPAFS
Panelists: Katherine K. Wallman, Office of Mgmt & Budget
Thomas Sawyer, US House of Representatives
Linda Gage, California State Finance Dept
Margo Anderson, U of Wisconsin-Milwaukee
Roderick J. Harrison, US Bur of the Census

231 M-Grand C/D
INTERACTIONS BETWEEN UNIVERSITY GRADUATE
PROGRAMS AND FOUR-YEAR COLLEGES—Invited Papers
*ASA Cmte on Career Development, Sec. on Stat. Educ., Sec. on Teaching of
Stat. in Hlth. Sci.*
Chair/Organizer: Rosemary A. Roberts, Bowdoin College
(10:35) FOUR-YEAR COLLEGES AS A SOURCE OF GOOD GRADUATE
STUDENTS. Thomas L. Moore, Grinnell College; Dean Isaacson, Iowa
State U
(11:00) RECRUITING A STATISTICIAN AT A FOUR-YEAR COLLEGE.
Gudmund Iversen, Swarthmore College; Philip J. Everson, Swarthmore
College
(11:25) PREPARING GRADUATE STUDENTS TO TEACH STATISTICS
AT A FOUR-YEAR COLLEGE. William I. Notz, Ohio State U; Ann R.
Cannon, Cornell College
(11:50) Disc: Lynne Billard, U of Georgia
(12:10) FLOOR DISCUSSION

232 H-Hunt
APPLIED ORDER-RESTRICTED INFERENCE—Invited I
ASA Council of Chapters
Organizer: Qing Liu, Food & Drug Admin
Chair: Roslyn A. Stone, U of Pittsburgh
(10:35) TESTING EQUALITY OF SURVIVAL CURVES UNDEF
CONSTRAINTS. Tim Wright, U of Missouri; Anura Abeyratne,
Co; Bahadur Singh, U of Missouri
(11:00) ORDERED INFERENCE IN CLINICAL TRIALS WITH
PLE ENDPOINTS. Dei-In Tang, Nathan Kline Inst for Psych Re:
(11:25) ORDER-RESTRICTED INFERENCE IN 2X2 TABLES V
HETEROGENEOUS ODDS RATIOS. Qing Liu, Food & Drug ⌐
(11:50) Disc: Jon H. Lemke, U of Iowa
(12:10) FLOOR DISCUSSION

233 M-Orange
PRACTICAL MARKOV CHAIN MONTE CARLO—Invit(
*Sec. on Bayesian Stat. Sci., ENAR, WNAR, IMS, Bio. Sec., Bus. & E
Sec., Stat. Comp. Sec.*
Organizer: James H. Albert, Bowling Green State U
Chair: Robert E. Kass, Carnegie-Mellon U
Panelists: Bradley P. Carlin, U of Minnesota
Andrew Gelman, Columbia U
Minghui Chen, Worcester Polytechnic Inst

234 H-Palos
MATCHING AND CONDITIONAL INDEPENDENCE: N
DEVELOPMENTS IN TESTING AND ESTIMATION—Invi
Papers
Bus. & Econ. Stat. Sec.
Organizer: James J. Heckman, U of Chicago
Chair: Robert Moffitt, Johns Hopkins U
(10:35) CONDITIONAL INDEPENDENCE RESTRICTIONS: 1
AND ESTIMATION. Oliver Linton, Yale U; Pedro Gozalo, Brov
(11:05) MATCHING AS AN ECONOMETRIC ESTIMATOR. H
Ichimura, U of Pittsburgh; Petra Todd, U of Pennsylvania
(11:35) ALTERNATIVE METHODS FOR EVALUATING SOCI/
PROGRAMS: THEORY AND EVIDENCE. James J. Heckman, l
Chicago
(12:05) FLOOR DISCUSSION

235 H-El Ca
JUDGEMENT IN OFFICIAL STATISTICS: HOW EXPLIC
SHOULD WE BE?—Invited Papers
Govt. Stat. Sec., Social Stat. Sec.
Chair/Organizer: Michael A. Stoto, Nat'l Academy of Sciences
Panelists: Jaime Marquez, Federal Reserve Board
Francisco J. Samaniego, U of California-Davis
Joseph Sedransk, Case Western Reserve U
Carl N. Morris, Harvard U

236 H-Huntir
LOST IN SPACE: ASSESSING MULTIVARIATE MISSIN(
DATA—Invited Papers
Sec. on Stat. Graph., ENAR, WNAR, Bio. Sec., Sec. on Hlth. Policy
Organizer: Dianne H. Cook, Iowa State U
Chair: Hal S. Stern, Iowa State U
(10:35) SENSITIVITY OF ANALYSES WITH MULTIVARIATE ?
DATA IN STUDIES OF THE ELDERLY. Robert J. Glynn, Brigh;
Women's Hosptial
(11:05) MISSING DATA IN INTERACTIVE HIGH-DIMENSIO
DATA VISUALIZATION. Deborah F. Swayne, Bellcore; Andrea:
Labs, Lucent Technologies
(11:35) CAN WE SEE WHAT ISN'T THERE? EXPLORING AN
KEEPING TRACK OF MISSINGS. Antony Unwin, Heike Hofn
Augsberg
(12:15) FLOOR DISCUSSION

○ = Theme session

ICSA 1997 Applied Statistics Symposium
May 30 - June 1, 1997
Rutgers University, New Jersey, USA
Title: Asymptotic Properties Of Self-Consistent Estimators of A Survival Function
by Qiqing Yu and George Y. C. Wong.
SUNY at Binghamton and Strang Cancer Prevention Center

ABSTRACT: The asymptotic properties of the nonparametric maximum likelihood estimator and other estimators of a joint distribution function F of a bivariate random vector X with right-censored data have been studied by several authors. Among others, an important assumption made in their studies is that X lives on a rectangle region $[0, a] \times [0, b]$ which can be observed. However, in many follow-up studies, $a = b = L$ is the length of the study period and X lives on a region larger than $[0, L] \times [0, L]$. Thus it is of interest to study whether the asymptotic results established by these authors are still valid without that restriction. In this direction, we established the strong consistency of self-consistent estimators of a discrete distribution function.

20th Annual
San Antonio
Breast Cancer
Symposium

December 3-6, 1997

July 30, 1997

George Y.C. Wong PhD
Strang Cancer Prevention Ctr
428 E 72 St
New York NY 10021

RE: *A dose-ranging study of indole-3-carbinol for breast cancer prevention.*

Dear Dr. Wong:

Your abstract referenced above has been accepted as a **poster** presentation (Program # 340) for the 20th Annual San Antonio Breast Cancer Symposium to be held December 3-6, 1997. Enclosed is a first draft copy of the program and instructions for a poster presentation. (Please review the instructions carefully, particularly the times for putting up and removing the posters, since our schedule is very tight.) The final program along with meeting registration and hotel information will be mailed to you soon -- you must still register for the meeting, even though your abstract has been accepted.

If for any reason your poster will not be presented, please notify Ms. Lois Dunnington as early as possible, by electronic mail (lois_dunnington@msmtp.idde.saci.org), by FAX (210-949-5009), or by phone (210-616-5912).

When you check in at the Registration Booth, your symposium materials will be given to you.

We look forward to your presentation at our symposium.

Sincerely,

GARY C. CHAMNESS, PhD
Chairman, Abstract Selection Committee
Medical Oncology

337 RETINOID-INDUCED GROWTH SUPPRESSION OF NORMAL HUMAN EPITHELIAL CELLS DOES NOT REQUIRE ACTIVATION OF RAR-DEPENDENT GENE TRANSCRIPTION. Yang, L-M., Ludes-Meyer, J., Munoz-Medellin, D., Kim, H-T., Reddy, P., Ostrowski, J., Reczek, P., and Brown, P. Division of Oncology, Dept. of Medicine, Univ. of Texas Health Science Center, San Antonio, TX, Bristol-Myers-Squibb, Albany, NY.

Retinoids inhibit the growth of breast cancer cells and are potential agents for cancer treatment and prevention. However, the mechanism by which retinoids prevent cancer is not known. The present studies investigated the mechanism by which naturally occuring and synthetic retinoids inhibit the growth of normal human mammary epithelial cells (HMECs). *All trans* retinoic acid (atRA) and 9cisRA both inhibited the growth of normal (184 and HMEC) and malignant (MCF7 and T47D) breast cells. We investigated whether retinoids inhibit normal breast growth by interfering with the cell cycle or inducing apoptosis. atRA treatment caused a cell cycle block (by increasing G0/G1 phase by 20% and decreasing S phase by 50%) and did not induce apoptosis. To explore the mechanism by which retinoids suppress cell growth, we correlated the growth inhibitory effects of retinoids with their ability to activate RAR-dependent transcription and transrepress AP-1-dependent transcription in breast cells. By measuring RAR-dependent transcription using a retinoid-responsive reporter and AP-1-dependent transcription using an AP-1 responsive reporter, we found that atRA and 9cisRA both activated RAR-dependent transcription in 184 normal breast cells and T47D breast cancer cells. atRA and 9cisRA also both inhibited AP-1 activity in T47D cells, while 9cisRA, but not atRA, inhibited AP-1 activity in 184 cells. Retinoid analogs which inhibit AP-1 without activating RAR were then used to determine whether inhibition of AP-1 without activation of RAR-dependent transcription was sufficient to inhibit breast cell growth. The growth of T47D and 184 was inhibited by these anti-AP-1 retinoids. These results suggest that RAR-dependent transcription is not required for retinoid-induced growth suppression of breast cells, which instead may be mediated by inhibition of AP-1. Such studies investigating the molecular mechanism by which retinoids inhibit breast cells growth may lead to the development of retinoid analogs for breast cancer prevention.

338 A multi-institutional study on the efficacy of prophylactic mastectomy in patients with Lobular Carcinoma in Situ (LCIS). Mackarem G*, Hughes KS, Berry D, Litten JB, Roche C, Veto J, Morris A, Turk P, Fraser H, Schnaper L, Friedman NB, Winer EP, Shafir M, Wanebo HJ, Capko D, Pories S, Khan S, Kroener J, Hawksworth K, Ting P, Barth R. Lahey Hitchcock Breast Center, Burlington, MA 01805.

Background: The efficacy of prophylactic mastectomy has not been adequately tested and yet women who carry BRCA1 and BRCA2 mutations are being offered this procedure. LCIS provides an established model of high risk for breast cancer. Published studies report that the risk of developing breast cancer in women with LCIS aproaches 33%. Our objective is to evaluate the efficacy of prophylactic mastectomy and to estimate lifetime risk reduction from this procedure.
Methods: Retrospective data on 493 patients with LCIS were collected from 14 institutions. Patients with the diagnosis of LCIS and no previous or synchronous DCIS or invasive cancer were eligible. 99 patients were treated with bilateral mastectomy (BMX), 74 patients were treated with ipsilateral mastectomy (IMX), and 320 patients were followed after initial biopsy (OBS). Ten year actuarial disease free survival (DFS) was calculated and compared for all groups, statistical significance between DFS was determined using the Mantel-Cox test.
Results: 17 patients developed an ipsilateral (IPSI), 12 a contralateral (CONT) and 1 a bilateral cancer, median time to recurrence was 63 months. One patient died with distant metastasis in the OBS group at 79 months.

Surgery	# pts.	FU(months)	IPSI	CONT	DFS	P
OBS	320	42	18 (6%)	12 (4%)	0.7694	
IMX	74	88	0	1 (1%)	0.9487	0.00001
BMX	99	75	0	-	1.0000	0.00001

16% of invasive recurrences were node positive.
Conclusions: (1)Prophylactic mastectomy markedly reduces the risk of cancer in patients with LCIS. (2) Prophylactic mastectomy has no impact on survival at 10 year. (3) This data can be useful when extrapolating results to patients with genetic predisposition.

339 Dietary Dehydroepiandrosterone (DHEA) Exhibits Strong Chemopreventive Activity But Minimal Therapeutic Activity In The MNU Induced Rat Mammary Model System.
Lubet, R.A.[1], Steele,V.E.[1], Kelloff, G.J.[1], Eto,I.[2], and Grubbs, C.J.[2]
1-Chemoprevention Branch NCI, Bethesda MD; 2- Dept. Of Nutrition Sciences, Univ. Of Alabama at Birmingham

Female Sprague-Dawley rats (50 day old) administered a single i.v. dose of MNU exhibit a high incidence and multiplicity of mammary tumors by 100 days of age . Prior studies have shown that DHEA (120, 600 and 2000 ppm in diet) is a highly effective chemopreventive agent in this model decreasing tumor multiplicity by 55, 90 and 98% respectively. DHEA doses ($\geq$600 ppm) causes striking hormonal changes in treated rats increasing levels of androgens and estrogens while simultaneously interfering with normal estrous cycling in rats. Interestingly DHEA induced proliferation and apparently differentiation in the breasts of treated rats. Morphologically the changes observed in DHEA treated rats appear similar to those that occur during pregnancy. When rats were treated with DHEA when their first palpable tumors arose (40-60 days post MNU) a decrease in the appearance of "new" late arising palpable tumors was observed. However DHEA had minimal effects on the continued growth of palpable lesions.

340 A DOSE-RANGING STUDY OF INDOLE-3-CARBINOL FOR BREAST CANCER PREVENTION. Wong GYC*, Bradlow L, Sepkovic D, Mehl S, Mailman J, Osborne MP, Strang Cancer Prevention Center, New York, NY, 10021

Sixty women at increased risk for breast cancer were enrolled in a placebo-controlled, double-blind dose-ranging chemoprevention study of indole-3-carbinol (I3C). Fifty-seven of these women with a mean age of 47 years (range 22-74) completed the study. Each women took a placebo capsule or an I3C capsule daily for a total of four weeks; none of the women experienced any significant toxicity effects. The urinary estrogen metabolite ratio of 2-hydroxyestrone to 16α-hydroxyestrone, as determined by an ELISA assay, served as the surrogate endpoint biomarker (SEB). Perturbation in the levels of SEB from baseline was comparable among women in the control (C) group and the 50, 100, 200 mg low dose (LD) group. Similarly, it was comparable among women in the 300, 400 mg high dose (HD) group. Regression analysis showed that peak relative change of SEB for women in the HD group was significantly greater than that for women in the C and LD groups by an amount that was inversely related to baseline ratio; the difference at the median baseline ratio was 0.48 with 95% confidence interval (0.30, 0.67). No other factors, such as age or menopausal status, were found to be significant in the regression analysis. The results in this study suggest that I3C at a minimum effective dose schedule of 300 mg per day is a promising chemopreventive agent for breast cancer prevention. A larger study to validate these results and to identify an optimal effective dose schedule of I3C for long-term breast cancer chemoprevention will be necessary. [Support: Tiger Foundation and U.S. Army Medical Research and Materiel Command under DAMD17-94-J-4332]

The Department of Defense
Breast Cancer Research Program Meeting

Era of Hope

The Ren   ce Hotel
Washi   n, DC
October 31   ember 4, 1997

FINAL PROGRAM

careful designed experiment should satisfy $P(X \in \mathcal{B}) = 0$. Thus this should not be a concern.

References

* Chang, M. N. (1990). Weak convergence of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.*, 18, 391-404.

* Chang, M. N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly censored data. *Ann. Statist.*, 15, 1536-1547.

* Ferguson, T. S. (1996). A course in large sample theory. *Chapman & Hall.* New York. 119-124.

* Gu, M. G. and Zhang, C-H. (1993). Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.*, 21, 611-624.

* Kim, M. Y., De Gruttola, V. G., and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*, 49, 13-22.

* Leiderman, P. H., Babu, D., Kagia, J., Kraemer, H. C. and Leiderman, G. F. (1973). African infant precocity and some social influences during the first year. *Nature*, 242, 247-249.

* Samuelsen, S. O. (1989). Asymptotic theory for non-parametric estimators from doubly censored data. *Scand. J. Statist.*, 16, 1-21.

* Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *JASA.*, 69, 169-173.

* Yu, Q. and Li, L. (1998). Asymptotic properties of self-consistent estimator with doubly-censored data. (Submitted for publication).

* Yu, Q. and Wong, G. Y. C. (1998). Technical report to "Modified GMLE with doubly-censored data". Department of Mathematical Sciences, SUNY at Binghamton.

501-W EFFECTS OF OMEGA-3 AND OMEGA-6 POLYUNSATURATED FATTY ACID(PUFA) DIETS ON INSULIN LIKE GROWTH FAC-TOR METABOLISM AND RODENT BREAST CANCER DEVELOPMENT
William T. Cave, Jr.
University of Rochester School of Medicine, Department of Internal Medicine, Rochester, NY

502-W EXPRESSION OF PEROXI-SOME PROLIFERATOR-ACTI-VATED RECEPTORS IN HUMAN AND RODENT MAMMARY TISSUES
Megan Lerner, Stan Lightfoot, Xiying Wu, Daniel Brackett, Alan Hollingsworth, and Jeff Gimble
Univ. of Oklahoma Health Sciences Center, Univ. of Oklahoma Inst. for Breast Health, Depts. of Surgery and Pathology; Veterans Affairs Medical Center; Oklahoma Medical Research Foundation, Immunobiology and Cancer Program, Oklahoma City, OK

503-W* BIOACTIVE LIPIDS IN BREAST CANCER
XiaoYan Zhang, Rina Das, and Marti Jett
Walter Reed Army Institute of Research, Washington, DC

504-W ROLE OF LIPOTROPE IN MAMMARY CARCINOGENESIS
Chung S. Park
North Dakota State University, Animal and Range Sciences Department, Fargo, ND

505-W DIETARY FAT ENHANCES THE EXPRESSION OF THE INDUC-IBLE PROSTAGLANDIN SYNTHASE (CYCLOOXYGENASE-2) IN HUMAN MAMMARY EPITHELIAL CELLS
Elizabeth A. Meade and Stephen M. Prescott
University of Utah, Human Molecular Biology and Genetics, and Huntsman Cancer Institute, Salt Lake City, UT

506-W C/EBP ISOFORM EXPRES-SION IN MOUSE MAMMARY TUMORS
Ronghua Yuan, Joy Lee, and James DeWille
The Ohio State University, Department of Veterinary Biosciences, Columbus, OH

506.1-W FATTY ACIDS AND BREAST CANCER RISK IN YOUNG WOMEN
J. L. Stanford, I. B. King, K. Malone, B. A. Blumenstein, and J. R. Daling

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA

X CLINICAL PRACTICES AND HEALTH CARE POLICY ISSUES 5:45 - 7:00 p.m.

507-X BREAST CANCER CARE: DOES PHYSICIAN GENDER MATTER?
Karen M. Freund, Risa B. Burns, Michelle Mancuso, Arlene Ash, and Mark A. Moskowitz
Boston Medical Center, Section of General Internal Medicine, Boston, MA

508-X EVALUATION OF NURSING CARE FOR WOMEN WITH NEWLY DIAGNOSED BREAST CANCER
Laurie Ritz, Laurel Decher, Brad Farrell, Karen Swenson, Lynne Schroeder, Mary Sladek, and Paul W. Sperduto
HealthSystem Minnesota, Minneapolis, MN

509-X INFORMATION FRAMING IN PHYSICIAN DESCRIPTION OF BREAST CANCER TREATMENT OPTIONS
K. R. Yabroff,[1,3] D. M. Seils,[1] L. E. Rubenstein,[1] C. Lerman,[2] N. J. Meropol,[6] B. O. Boekeloo,[4] D. M. Brown,[5] C. Weaver,[7] and K. A. Schulman[1]
Georgetown Univ. Med. Ctr., [1]Clinical Econ. Res. Unit and [2]Lombardi Cancer Ctr.; [3]MEDTAP, Intl.; [4]George Washington Univ.; [5]Georgetown Univ., Washington, DC; [6]Roswell Park Cancer Inst., Buffalo, NY; [7]Response Oncoloy, Inc., Memphis, TN

510-X DEVELOPMENT OF A NOVEL CLINICAL CURRICULUM TO EVALUATE BREAST CANCER PATIENT OUTCOMES AND DETERMINE PRIORITIES FOR HEALTH CARE REFORM
Garrett A. Smith
University of California, San Francisco, CA

511-X A CLINICAL DECISION SUPPORT SYSTEM FOR BREAST CANCER
Harry Burke, Philip Goodman, and Albert Hoang
New York Medical College, Valhalla, NY; University of Nevada School of Medicine, Reno, NV

512-X WEB-BASED BREAST CAN-CER DATA QUERY AND CLINICAL TRIAL MATCHING SYSTEMS
Barry A. Gordon and Alan Houser
Public Health Institute, C/NET Solutions, Berkeley, CA

513-X STATISTICAL METHODS FOR BREAST CANCER FOLLOW-UP STUDIES INVOLVING INTER-VAL CENSORING
George Y.C. Wong[1] and Qiqing Yu[2]
[1]Strang Cancer Prevention Center, New York, NY; [2]State University of New York at Binghamton, Binghamton, NY

514-X NEURAL NETWORKS FOR BREAST CANCER PROGNOSIS
Jonathan Buckley, Jan VanTornout, Leslie Bernstein, Michael Press, and Kerry Flom
University of Southern California, Department of Preventive Medicine, Los Angeles, CA

515-X A SUBACUTE CARE INTER-VENTION FOR SHORT-STAY BREAST CANCER SURGERY
Gwen Wyatt
Michigan State University College of Nursing, East Lansing, MI

516-X UNDERUSE OF SURVEIL-LANCE MAMMOGRAPHY AFTER BREAST CANCER TREATMENT IN A MEDICARE POPULATION
Marilyn M. Schapira, Timothy L. McAuliffe, and Ann B. Nattinger
Medical College of Wisconsin, Divisions of General Internal Medicine and Biostatistics, Milwaukee, WI

517-X ACCURACY OF INPATIENT MEDICARE CLAIMS FOR BREAST CANCER THERAPY DETERMINA-TION
Craig A. Beam and Ann B. Nattinger
Medical College of Wisconsin, Divisions of General Internal Medicine and Family and Community Medicine, Milwaukee, WI

Y INFRASTRUCTURE (TISSUE BANKS AND REGISTRIES) 5:45 - 7:00 p.m.

518-Y CAROLINA MAMMOGRA-PHY REGISTRY. DEVELOPMENT AND EARLY RESULTS
Bonnie C. Yankaskas,[1] Hope S. Carlson,[1] Kara T. Gasink,[1] Jennifer L. David,[1] Susan A. Maygarden,[2] and Tim E. Aldrich[3]

ASYMPTOTIC PROPERTIES OF SELF-CONSISTENT ESTIMATORS WITH MIXED INTERVAL-CENSORED DATA

By Qiqing Yu, [1], George Y. C. Wong [1] and Linxiong Li [2]

Department of Mathematical Sciences, SUNY, Binghamton, NY 13902, USA
Strang Cancer Prevention Center, 428 E 72nd Street, NY 10021, USA
and
Department of Mathematics, University of New Orleans, LA 70148, USA

Short title: Mixture interval censorship model.
VERSION: 8/28/99.
AMS 1991 subject classification: Primary 62 G05; Secondary 62 G20.

Abstract

Mixed interval-censored (MIC) data consist of n pairs of observations (L_1, R_1), ..., (L_n, R_n), where $-\infty \leq L_i \leq R_i \leq \infty$ for all i, $L_k = R_k$ and $0 < L_j < R_j < \infty$ for at least one k and one j. The survival time X_i is only known to lie between L_i and R_i, $i = 1, 2, \ldots, n$. Peto (1973) and Turnbull (1976) obtained, respectively, the generalized MLE (GMLE) and the self-consistent estimator (SCE) of the distribution function of X with MIC data. In this paper, we introduce a model for MIC data and establish strong consistency, asymptotic normality and asymptotic efficiency of the SCE and GMLE with MIC data under this model with mild conditions.

Key words and phrases: Asymptotic normality, generalized maximum likelihood estimator, mixture distribution, strong consistency.

1. Introduction

Interval censoring refers to a situation in which, X, the time to occurrence of an event of interest is only known to lie in a half-open and half-closed time interval $(L, R]$, where the pair (L, R) is an extended random vector such that $-\infty \leq L < X \leq R \leq \infty$. Interval-censored (IC) data may occur in medical follow-up studies when each patient had several visits and the event of interest was only known to take place either before the first visit, between two consecutive visits, or after the last one. Thus an IC data set may consist of strictly interval-censored (SIC) observations (*i.e.*, $0 < L < R < \infty$), and right-censored ($R = \infty$) and/or left-censored ($L = -\infty$) observations. Examples of IC data can be found in cancer research and AIDS studies (see, *e.g.*, Finkelstein and Wolfe, 1985).

Case 1 data (or current status data, see Ayer *et al.*, 1955) is a special case of IC data when each patient had only one visit. Observations in a case 1 data set are either left-censored or right-censored. Doubly-censored data (see Chang and Yang, 1987) consist of case 1 data and uncensored observations. It is clear that neither case 1 data nor IC data contain uncensored observations. Furthermore, doubly-censored data do not contain SIC observations. A data set may be a mixture of uncensored observations and IC data which contain SIC observations. We call such data *mixed interval-censored (MIC) data.*

MIC data arise in clinical follow-up studies. In a cancer follow-up study, a patient whose tumor marker value (for instance, CA 125 in ovarian cancer) is consistently on the high (or low) end of the normal range in repeated testing is usually monitored very closely for possible relapse. If such a patient should relapse, then time to clinical relapse can often be accurately determined, and an uncensored observation is obtained. However, if a patient is not under close surveillance, and would seek help only after some tangible symptoms of the disease have appeared, then time to relapse most likely has to be specified to be within the dates of two successive clinical visits.

Another situation in which MIC data can occur is in the usual right-censored survival analysis where actual dates of events are not recorded, or missing, for a subset of the study population, and can be established only to within specified intervals. An example from the Framingham Heart Study was presented by Odell *et al.* (1992). In this large-scale longitudinal heart disease study, times of occurrence of coronary heart disease were recorded for almost every participant. However, time of first occurrence of the coronary heart disease subcategory angina pectoris was only recorded for about 20% of the participants who suffered from angina pectoris, and may be specified only as between two clinical visits, several years apart, for the other participants.

For censored data, Peto (1973) proposed a Newton-Raphson algorithm to obtain the generalized MLE (GMLE) of the cumulative distribution function (cdf), F. Turnbull (1976) obtained a self-consistent estimator (SCE) of the cdf via an EM-algorithm. A detailed discussion of more efficient algorithms for obtaining the GMLE is given in Wellner and Zhan (1997).

For IC data, Groeneboom and Wellner (1992) formulated the case 2 model; Wellner (1995) formulated a case k model, where $k \geq 1$; Schick and Yu (1999) modified Wellner's case k model by further assuming that k, the number of visits by a patient in a follow-up study, is a random integer and the observation (L, R) is a mixture of various case k models.

Various asymptotic distribution results of the GMLE have been obtained for censored data. For case 1 model the GMLE is asymptotically normally distributed (a.n.) and the convergence rate is $n^{1/2}$ if the underlying censoring distribution is discrete (Yu *et al.*, 1998b), but the GMLE is not a.n. and the convergence rate is $n^{1/3}$ if cdfs have positive derivatives (Groeneboom and Wellner, 1992). For case 2 model the GMLE is a.n. with rate $n^{1/2}$ if the censoring vector takes on finitely many values (Yu *et al.*, 1998c), and Groeneboom and Wellner's (1992) conjecture that under certain smoothness conditions the GMLE has a pointwise convergence rate of $(n \ln n)^{1/3}$. For more recent development on the latter conjecture, we refer to Groeneboom (1996) and Van De Geer (1996) .

For MIC data, several models have been proposed, and the asymptotic properties of the GMLE have been investigated under the assumptions that either the censoring vector takes on finitely many values (see Petroni and Wolfe, 1994, and Yu *et al.* 1998a), or the censoring and survival distributions are strictly increasing and continuous, and they have "positive separation" (see Huang (1999)).

In this paper, we shall use the model in Yu *et al.* (1998a) to establish asymptotic properties of the GMLE based on MIC data under the assumption that all underlying distributions are arbitrary with some mild conditions. Since a GMLE is also an SCE (but an SCE may not be a GMLE; see Yu *et al.*, 1998a),

2

and our proofs basically use the properties of SCEs, we shall focus on the asymptotic properties of SCEs for MIC data. The main results are given in Section 2. The consistency result is proved in Section 3 and the asymptotic normality result is proved in Section 4. Some detailed proofs of lemmas in Sections 3 and 4 are relegated to Appendices A and B.

2. Main Results

We introduce a *mixture interval censorship* model, a mixture of an interval censorship model and a right censorship model, to characterize MIC data. Assume that the observed pair (L, R) is generated by a two-stage experiment. Let (T, U, V) be a random censoring vector and $\mathcal{K}$ a random integer taking values 0 and 2. Assume that X and $(\mathcal{K}, T, U, V)$ are independent. In the first stage, a value of $\mathcal{K}$ is selected, then (L, R) corresponds to the observation from a right censorship model if $\mathcal{K} = 0$ and from a case 2 model if $\mathcal{K} = 2$, i.e.,

$$(L, R) = \begin{cases} (X, X)\mathbf{1}_{(X \leq T)} + (T, \infty)\mathbf{1}_{(X > T)} & \text{if } \mathcal{K} = 0, \\ (-\infty, U)\mathbf{1}_{(X \leq U)} + (U, V)\mathbf{1}_{(U < X \leq V)} + (V, \infty)\mathbf{1}_{(X > V)} & \text{if } \mathcal{K} = 2, \end{cases} \tag{2.1}$$

where $\mathbf{1}_{(A)}$ is the indicator function of the set A. It is known that in order to estimate F, we only need to observe (L, R) (see Peto (1973)). Thus, in our model, X, $\mathcal{K}$, U, V and T may not be observed. Let $\pi_k = P(\mathcal{K} = k) > 0$, $k = 0, 2$, and $\pi_0 + \pi_2 = 1$. Denote $(L_1, R_1), \ldots, (L_n, R_n)$ a random sample from (L, R).

Suppose that X, (L, R), (U, V), T, U and V have cdfs F, Q, G, G_T, G_U and G_V, respectively. Define $\tau_o = \sup\{x : F(x) = 0\}$, $\tau_v = \sup\{x : G_V(x) < 1\}$, $\tau_t = \sup\{x : G_T(x) < 1\}$ and $\tau = \inf\{x : F(x) = 1 \text{ or } G_T(x) = 1\}$. Let $\Theta = \{h : h$ is a nondecreasing function from $[-\infty, \infty]$ to $[0, 1]$ such that $h(-\infty) = 0$ and $h(\infty) = 1\}$. Each solution H_n of the equation

$$H_n(x) = \int_{l < x < r} \frac{H_n(x) - H_n(l)}{H_n(r) - H_n(l)} dQ_n(l, r) + \int_{r \leq x} dQ_n(l, r), \quad H_n \in \Theta \tag{2.2}$$

is an SCE of F (Li, Watkins and Yu, 1997), where Q_n is the empirical version of Q.

Theorem 2.1. *Let H_n be a solution of (2.2). Suppose that*

(AS1)   *(a)* $\tau_v \leq \tau_t$, *and (b) if* $F(\tau_t -) < 1$ *then* $P\{T \text{ or } V = \tau_t\} > 0$.

Then $\lim_{n \to \infty} \sup_{x \geq 0} |H_n(x) - F(x)| = 0$ *a.s. if* $F(\tau) = 1$, *and* $\lim_{n \to \infty} \sup_{x \leq \tau} |H_n(x) - F(x)| = 0$ *a.s..*

Remark 1. *A counterexample similar to that in Schick and Yu (1999) can be constructed to show that the GMLE is not consistent if AS1.b is deleted from our Theorem 2.1.*

In clinical follow-ups, a study typically lasts for a certain period of time. Thus it is often true that $F(\tau-) < 1$. In this regard, Gentleman and Geyer (1994, Theorem 2) claimed a vague convergence result, and Huang (1996, Theorem 3.1) claimed a uniform strong consistency result for IC data or case 1 data. Schick and Yu (1999) showed that both theorems as stated are false and can be corrected by adding assumption AS1.b to their theorems.

It is well known (see Peto, 1973) that a GMLE $\hat{F}_n(t)$ is not uniquely determined for $t \in (L_i, R_j)$ if $L_i < R_j$, $(L_i, R_j) \cap \{L_1, \ldots, L_n, R_1, \ldots, R_n\} = \emptyset$ and $\mu_{\hat{F}_n}((L_i, R_j]) > 0$. For the convenience of our proof of normality, we restrict our attention to the following SCEs:

$$H_n \text{ is right continuous, } H_n(\infty) = 1 \text{ and } \mathcal{S}_{H_n} \subset \{R_1, \ldots, R_n\}. \tag{2.3}$$

Under convention (2.3) the GMLE $\hat{F}$ is uniquely determined. However there are still SCEs that satisfy (2.3) but are not the GMLE. A point x is called a *support point of* a function f if there exists a sequence of points $x_k \to x$ such that $|f(x_k) - f(x)| > 0$. Denote $\mathcal{S}_f$ the set of all support points of f.

Theorem 2.2. *Let H_n satisfies (2.2) and (2.3). Suppose that AS1 holds and*

(AS2) $F(\tau) > 0$ *and* $(\mathcal{S}_{G_U} \cup \mathcal{S}_{G_V}) \subset \mathcal{S}_F$.

Then for $x \leq \tau$, $\sqrt{n}(H_n(x) - F(x))$ converges in distribution to a normal variate.

AS1 and AS2 are much weaker than the assumptions made in Petroni and Wolfe (1994), Yu *et al.* (1998a) and Huang (1999).

Remark 2. *In a follow-up study, each patient has $\mathcal{N}$ visits, where $\mathcal{N} \geq 1$ is a random integer (rather than assuming that each patient has exactly 2 visits ($\mathcal{N} \equiv 2$) as in the case 2 model). The inspection times are $Y_1 < \cdots < Y_{\mathcal{N}}$. It is reasonable to assume that X and $(\mathcal{N}, \{Y_i : i \geq 1\})$ are independent. Then, on the event $\{\mathcal{N} = k\}$, modify (U, V) in (2.1) as*

$$(U, V) = (Y_1, Y_2)\mathbf{1}_{(X \leq Y_1)} + (Y_{k-1}, Y_k)\mathbf{1}_{(X > Y_k)} + \sum_{i=2}^{k} (Y_{i-1}, Y_i)\mathbf{1}_{(Y_{i-1} < X \leq Y_i)}, \tag{2.5}$$

where $Y_0 = 0$. Thus, a more realistic model for MIC data is the model of a mixture of a right censorship model and a modified case 2 model where (U, V) is specified by (2.5), instead of assuming that X and (U, V) are independent. This model includes our model (2.1) (in which $\mathcal{N} = 2$ with probability one) as well as Huang's model (in which $\mathcal{N}$ is a fixed positive integer and $T = \infty$). It is reasonable to assume that $\mathcal{N}$, the number of visits, is bounded. In such a model the proofs of Theorems 2.1 and 2.2 are similar to the proofs given in Sections 3 and 4. Thus it suffices to study model (2.1).

3. Strong Consistency

We shall prove Theorem 2.1. To this end, we first state two preliminary results.

<u>**Theorem 3.1.**</u> *Suppose that $F \in \Theta$, F is right continuous and H is a solution of*

$$H(x) = \int_{l < x < r} \frac{H(x) - H(l)}{H(r) - H(l)} dQ(l, r) + \int_{r \leq x} dQ(l, r), \quad H \in \Theta. \tag{3.1}$$

Then $H(x) = F(x)$ for all $x \leq \tau$ if AS1 holds; and $H(x) = F(x)$ for all $x < \tau_t$ if
(AS3) $F(\tau_t) < 1$, $\tau_v \leq \tau_t$ and $F = F(\tau_t)$ on $[x_o, \infty)$, where $x_o < \tau_t$.

In (3.1), if $H(x) = H(r) = H(l)$, then we encounter $\frac{0}{0}$ in the integrand. Hereafter, define $\frac{0}{0} = 1$ and $\frac{0}{0} \cdot 0 = 0$. If F satisfies AS3, it can viewed as the cdf of an extended random variable X which equals ∞ with positive probability.

<u>**Proposition 3.2.**</u> *Suppose that $\{f_n\}_{n \geq 1}$ is a sequence of monotone functions on an interval $[a, b)$ and $f(x)$ is a bounded monotone and right continuous function on the same interval. If $\lim_{n \to \infty} f_n(x) = f(x) \; \forall \; x \in [a, b)$ and $\lim_{n \to \infty} f_n(x-) = f(x-) \; \forall \; x \in (a, b]$, then $\lim_{n \to \infty} \sup_{x \in [a,b)} |f_n(x) - f(x)| = 0$.*

We shall present the proof of Theorem 3.1 after we prove Theorem 2.1. We omit the proof of Proposition 3.2 as it is similar to Lemma 3 of Yu and Li (1994).

Proof of Theorem 2.1. Let Ω be the event $\{\lim_{n \to \infty} Q_n(l, r) = Q(l, r) \; \forall \; l < r\}$. For each $\omega \in \Omega$, let H_n be a solution of (2.2). We shall prove the theorem in 2 steps.

Step 1 ($\lim_{n \to \infty} H_n(x) = F(x)$ and $\lim_{n \to \infty} H_n(x-) = F(x-) \; \forall \; x \leq \tau$). Since $\{H_n\}_{n \geq 1}$ is bounded and monotone, for each subsequence of natural numbers, by Helly's selection theorem, there exists a further subsequence, say $\{n_k\}$, such that $\lim_{n_k \to \infty} H_{n_k}(x) = H(x)$ and $\lim_{n_k \to \infty} H_{n_k}(x-) = H^*(x)$ pointwisely for some H and $H^* \in \Theta$, respectively. Thus it suffices to show that $H(x) = F(x)$ and $H^*(x) = F(x-)$ for all $x \leq \tau$.

Since Q_n converges uniformly to Q, and H_n satisfies (2.2), by the bounded convergence theorem (BCT) H satisfies (3.1) and H^* satisfies a similar equation like (3.1). Theorem 3.1 yield the first desired equation $H(x) = F(x)$ on $(-\infty, \tau]$.

By AS1.a, $r > \tau \Rightarrow r = \infty$ and thus $H(r) = F(r) = 1$ as $H \in \Theta$. Then equation (3.1) and its analog for H^* yield

$$F(x-) = \int_{l < x \leq r} \frac{F(x-) - F(l)}{F(r) - F(l)} dQ(l, r) + \int_{r < x} dQ(l, r) \quad (\text{as } \int_{l < x < r} + \int_{r \leq x} = \int_{l < x \leq r} + \int_{r < x}),$$

$$H^*(x) = \int_{l < x \leq r} \frac{H^*(x) - F(l)}{F(r) - F(l)} dQ(l, r) + \int_{r < x} dQ(l, r),$$

as $H = F$ on $(-\infty, \tau] \cup \{\infty\}$. The latter two equations yield

$$H^*(x) - F(x-) = (H^*(x) - F(x-))c(x), \quad \text{where } c(x) = \int_{l < x \leq r} \frac{1}{F(r) - F(l)} dQ(l, r). \tag{3.2}$$

4

By AS1, $c(x) = \begin{cases} 1 - \pi_0 P(T > x) < 1 & \text{if } x < \tau, \\ 1 - P(L = \tau) < 1 & \text{if } x = \tau \text{ and } F(\tau-) < 1. \end{cases}$ It follows from equation (3.2) and $c(x) < 1$ that $H^*(\tau) = F(\tau-)$ if $F(\tau-) < 1$, and $H^*(x) = F(x-) \ \forall \ x < \tau$. In order to show that $H^*(\tau) = F(\tau-)$ if $F(\tau-) = 1$, let $x_k \uparrow \tau$. Note $H_n(x_k) \le H_n(\tau-) \le 1$. It yields $H(x_k) \le H^*(\tau) \le 1$. Now $\lim_{k\to\infty} H(x_k) = \lim_{k\to\infty} F(x_k) = 1$. Thus $H^*(\tau) = 1 = F(\tau-)$.

Step 2 (conclusion). By step 1 the sequence $\{H_n\}_{n\ge 1}$ and F satisfy all the conditions for $\{f_n\}_{n\ge 1}$ and f in Proposition 3.2, respectively, where $(a,b) = (-\infty, \tau)$. By Proposition 3.2, $\lim_{n\to\infty} \sup_{x\le\tau} |H_n(x) - F(x)| = 0 \ \forall \ \omega \in \Omega$. Since $P\{\Omega\} = 1$, Theorem 2.1 follows. □

The solution $H(x)$ to (3.1) is unique for $x < \tau_t$ if AS3 holds by Thereom 3.1, but Theorem 2.1 is false if only AS3 holds, as $\int_{l < \tau_t \le r} \frac{1}{F(r)-F(l)} dQ(l,r) = 1$ if $P(T < \tau_t) = 1$ and $P(V < \tau_t) = 1$. The rest of the section is devoted to prove Theorem 3.1.

The theorem is trivially true if $F(\tau) = 0$, so without loss of generality (WLOG), we can assume $F(\tau) > 0$. The outline of the proof is as follows. We first define a functional $\psi(h)$ for $h \in \Theta$. We then show that $h = F$ uniquely maximizes $\psi(h)$ for $h \in \Theta$ (Lemma 3.3) and that each solution H of (3.1) in Θ is a maximum point of $\psi(\cdot)$. Thus H must equal F. To this end, some notations and lemmas are needed.

Verify that there are at most countably many intervals (y, z) such that (1) $y < z$ and $y \le \tau$, (2) $F(y) = F(z-)$, and (3) $y, z \in S_F$. Let $\mu(x) = [F(x) + G_U(x) + G_V(x) + G_T(x)]/4$. For $i \ge 1$, denote D_i the collection of intervals (y, z) satisfying (1), (2), (3) above and $\mu(z-) - \mu(y) \ge 1/i$, then D_i contains finitely many intervals since $\mu(\cdot)$ is a cdf. Thus $\cup_i D_i$, the collection of all such intervals, is countable. Denote D_i^e the set of left endpoints of intervals in D_i.

For $\alpha = 1, 2, ...,$ denote $B_{\alpha,1}$ the collection of all possible $j2^{-\alpha} \times 100$ percentiles of the distribution μ $(1 \le j \le 2^\alpha)$ which are contained in $(-\infty, \tau]$. Note that for each j such that $j2^{-\alpha} \le \mu(\tau)$ the corresponding percentile is given by $y = \sup\{x : \mu(x) < j2^{-\alpha}\}$. Let $B_\alpha = (B_{\alpha,1} \cup D_\alpha^e) \cup \{\tau\}$ and denote $b_1 < \cdots < b_\beta = \tau$ to be the elements of B_α. Verify that

$$\mu(b_i-) - \mu(b_{i-1}) \le 2^{-\alpha}, \ i = 2, ..., \beta. \tag{3.3}$$

Define $b_{1*} = b_1$ and $b_{i*} = \sup\{x : x \le b_i, \ F(x) = F(b_{i-1})\}$, $i = 2, ..., \beta$. Moreover, if $\tau < \infty$, then denote $b_{\beta+1*} = \tau$ and $b_{\beta+1} = \infty$. For $b_i, b_j \in B_\alpha$, define

$$T_\alpha = \sum_{i=1}^{\beta+1} b_i \mathbf{1}_{(T \in [b_i, b_{i+1}))}, \quad \lfloor b_{i*}, b_i \rfloor = \begin{cases} [b_{i*}, b_i] & \text{if } b_{i*} > b_{i-1}, \\ (b_{i*}, b_i] & \text{if } b_{i*} = b_{i-1}, \end{cases} \text{ and}$$

$$(U_\alpha, V_\alpha) = (b_i, b_j) \text{ if } b_i \le U < b_{i+1}, \ b_{j-1} < V \le b_j, \ i \le j \le \beta. \tag{3.4}$$

Then $P\{X \in (b_{i-1}, b_i]\} = P\{X \in \lfloor b_{i*}, b_i \rfloor\}$ as $P\{X \in (b_{i-1}, b_{i*})\} = 0$. Define an interval

$$I_\alpha = \begin{cases} (-\infty, b_i] & \text{if } \mathcal{K} = 2 \text{ and } X \le b_i = U_\alpha, \\ (b_i, b_j] & \text{if } \mathcal{K} = 2, \ X \in (b_i, b_j] \text{ and } (U_\alpha, V_\alpha) = (b_i, b_j), \\ (b_i, \infty] & \text{if } X > b_i \text{ and either } \mathcal{K} = 2 \text{ and } V_\alpha = b_i \text{ or } \mathcal{K} = 0 \text{ and } T_\alpha = b_i, \\ \lfloor b_{i*}, b_i \rfloor & \text{if } X \in \lfloor b_{i*}, b_i \rfloor, \ \mathcal{K} = 0 \text{ and } T_\alpha \ge b_i. \end{cases} \tag{3.5}$$

Then the number of distinct realizations $I_{\alpha,h}$ of the random interval I_α is finite. Denote the joint cdf of (U_α, V_α) by G_α and the cdf of T_α by G_{T_α}. Let L^α and R^α be the endpoints of the interval I_α, $Q_\alpha(l, r, k)$ the joint cdf of $(L^\alpha, R^\alpha, \mathcal{K})$, and $q_{\alpha,h,k} = P(I_\alpha = I_{\alpha,h}, \mathcal{K} = k)$. Abusing notations, let $Q(l, r, k)$ be the joint cdf of $(L, R, \mathcal{K})$. Thus $Q(l, r)$ can be viewed as the marginal cdf of (L, R).

For $H \in \Theta$, define μ_H to be the measure induced by H and

$$\psi_\alpha(H) = E[\ln(\mu_H(I_\alpha)/\mu_F(I_\alpha))] \ \left(= \sum_{h,k} q_{\alpha,h,k} \ln[\mu_H(I_{\alpha,h})/\mu_F(I_{\alpha,h})]\right). \tag{3.6}$$

Here we interpret $\ln 0 = -\infty$, $0\ln 0 = 0$ and $0\ln\infty = 0$. It is obvious by construction [see (3.3), (3.4) and (3.5)] and by AS1.a that the measures dG_α, dG_{T_α} and dQ_α converge setwisely to dG, dG_T and dQ, respectively. We call $\psi(H)$ a *limit* of $\{\psi_\alpha(H), \alpha \ge 1\}$ if a subsequence of $\{\psi_\alpha(H)\}$ converges to $\psi(H)$, where $\psi(H)$ may be ∞.

The proofs of the following 2 lemmas are given in Appendix A.

Lemma 3.3. *Suppose that $H \in \Theta$ and either AS1 or AS3 holds. Let $\psi(H)$ be a limit of $\{\psi_\alpha(H)\}$. Then (1) $\psi(H) = 0$ if and only if $H(x) = F(x)$ for all $x < \tau$, and $H(\tau_t) = F(\tau_t)$ in the case $F(\tau_t-) < 1$ and $P(T$ or $V = \tau_t) > 0$; (2) $\psi(H) \leq 0$.*

A real number $x \in [\tau_o, \tau]$ is called a *left point of increase of* $F \in \Theta$ if $F(x) - F(x - \epsilon) > 0$ for each $\epsilon > 0$. Let $\mathcal{L}_F$ be the set of all left points of increase of F.

Denote
$$\gamma_H(a, b) = \frac{P(X \in (a, b], X \leq T, \mathcal{K} = 0)}{H(b) - H(a)}.$$

Lemma 3.4. *Suppose that H is a solution of (3.1), AS1 or AS3 holds, and $b \in \mathcal{L}_F$. Then*
(E.1) $\int \frac{\mathbf{1}_{(l \leq \tau_t < r)}}{H(r) - H(l)} dQ(l, r) = 1$ *if* $F(\tau) < 1$;

(E.2) $\gamma_H(a, b) \leq 1$ *for each* $a < b$; *(E.3)* $\int_{l<r} \frac{\mathbf{1}_{(b \in (l, r])}}{H(r) - H(l)} dQ(l, r) + \lim_{a \uparrow b} \gamma_H(a, b) - 1 = 0$.

Proof of Theorem 3.1. Let H be a solution of (3.1). We shall assume that $H(x) \neq F(x)$ for some $x \leq \tau$ but AS1 holds, or $H(x) \neq F(x)$ for some $x < \tau$ but AS3 holds; and show that it leads to a contradiction.

Let $\psi(H)$ be a limit of $\psi_\alpha(H)$. WLOG, assume $\lim_{\alpha \to \infty} \psi_\alpha(H) = \psi(H)$. Since $H \neq F$ for some $t_0 \leq \tau$, $\psi(F) = 0 > \psi(H)$ by Lemma 3.3. Therefore, there exists an integer α_1 such that $\psi_\alpha(F) > \psi_\alpha(H) + \delta$, for all $a \geq \alpha_1$, where $\delta = -\psi(H)/2 > 0$. For each $\alpha \geq \alpha_1$, let $p_i = \mu_F([b_{i*}, b_i])$, $i = 1, ..., \beta$, and $p_{\beta+1} = 1 - F(\tau)$. It is seen that b_i, β, and p_i all are functions of α. Then, for $\alpha \geq \alpha_1$, the above inequality yields

$$
\begin{aligned}
\delta &\leq -\psi_\alpha(H) + \psi_\alpha(F) \\
&= \lim_{u \downarrow 0} \frac{\frac{1}{1+u} \psi_\alpha(H) + \frac{u}{1+u} \psi_\alpha(F) - \psi_\alpha(H)}{u} \\
&\leq \lim_{u \downarrow 0} \frac{\psi_\alpha(\frac{1}{1+u} H + \frac{u}{1+u} F) - \psi_\alpha(H)}{u} \quad \text{(since } -\ln(\cdot) \text{ and hence } -\psi_\alpha(\cdot) \text{ is convex)} \\
&= \lim_{u \downarrow 0} \frac{\sum_{j,k} q_{\alpha,j,k} \ln \frac{\mu_{\frac{H+uF}{1+u}}(I_{\alpha,j})}{\mu_F(I_{\alpha,j})} - \sum_{j,k} q_{\alpha,j,k} \ln \frac{\mu_H(I_{\alpha,j})}{\mu_F(I_{\alpha,j})}}{u} \quad \text{(by (3.6))} \\
&= \lim_{u \downarrow 0} \frac{\sum_{j,k} q_{\alpha,j,k} [\ln(1 + \frac{u \mu_F(I_{\alpha,j})}{\mu_H(I_{\alpha,j})}) - \ln(1 + u)]}{u} \quad \text{(as } \frac{H+uF}{1+u} = \frac{1}{1+u} H + \frac{u}{1+u} F) \\
&= \sum_{j,k} q_{\alpha,j,k} \frac{\mu_F(I_{\alpha,j})}{\mu_H(I_{\alpha,j})} - 1 \\
&= \int_{l<r \text{ and } k=2, \text{ or } r=\infty} \frac{F(r) - F(l)}{H(r) - H(l)} dQ_\alpha(l, r, 2) + \sum_{i=1}^{\beta} p_i \frac{q_{\alpha,j_i,0}}{\mu_H([b_{i*}, b_i])} - 1, \quad (3.7)
\end{aligned}
$$

where j_i is such that $I_{\alpha,j_i} = [b_{i*}, b_i]$, $i = 1, ..., \beta$. Let $h_1(l, r) = \frac{F(r) - F(l)}{H(r) - H(l)}$ and $h_2(b_{i*}, b_i) = \frac{q_{\alpha,j_i,0}}{\mu_H([b_{i*}, b_i])}$. By (E.1), (E.2) and (E.3) in Lemma 3.4, $\int_{l<r} \frac{\mathbf{1}_{(b_j \in (l,r])}}{H(r) - H(l)} dQ(l, r) \leq 1$, thus

$$
\begin{aligned}
\infty &> \lim_{\alpha \to \infty} \sum_{j=1}^{\beta+1} p_j \int_{l<r} \frac{\mathbf{1}_{(b_j \in (l,r])}}{H(r) - H(l)} dQ(l, r) \quad \text{(by the BCT)} \\
&\geq \int_{l<r} \frac{\lim_{\alpha \to \infty} \sum_{j=1}^{\beta+1} p_j \mathbf{1}_{(b_j \in (l,r])}}{H(r) - H(l)} dQ(l, r) \quad \text{(by Fatou's lemma)} \\
&= \int_{l<r} h_1(l, r) dQ(l, r). \quad (3.8)
\end{aligned}
$$

Since h_1 is a nonnegative measurable function, (3.8) implies that it is integrable. Since

$$q_{\alpha,j_i,0} \leq P(X \in [b_{i*}, b_i], X \leq T_\alpha, \mathcal{K} = 0) \quad (3.9)$$

6

by the definition of U_α, V_α, T_α and I_α [see (3.4) and (3.5)], (E.2) and (3.9) imply that $|h_2(b_{i*}, b_i)| \leq 1$, and thus $\sum_{i=1}^\beta p_i h_2(b_{i*}, b_i)$ converges by the BCT as $\alpha \to \infty$. Then

$$0 < \delta \leq \text{expression (3.7)}$$

$$\leq \varliminf_{\alpha \to \infty} \left[\int_{l < r \text{ and } k=2, \text{ or } r=\infty} h_1(l,r) dQ_\alpha(l,r,k) + \sum_{i=1}^\beta p_i h_2(b_{i*}, b_i) - 1 \right]$$

$$= \int_{l<r} h_1(l,r) dQ(l,r) + \lim_{\alpha \to \infty} \sum_{i=1}^\beta p_i h_2(b_{i*}, b_i) - 1 \quad \text{(since } dQ_\alpha \to dQ \text{ setwisely)}$$

$$\leq \int_{l<r} h_1(l,r) dQ(l,r) + \lim_{\alpha \to \infty} \sum_{i=1}^\beta p_i \gamma_H(b_{i*}, b_i) - 1 \qquad \text{(by (3.9))}$$

$$\leq \varliminf_{\alpha \to \infty} \sum_{i=1}^{\beta+1} p_i \int \frac{\mathbf{1}_{(b_i \in (l,r])}}{H(r) - H(l)} dQ(l,r) + \varlimsup_{\alpha \to \infty} \sum_{i=1}^\beta p_i \gamma_H(b_{i*}, b_i) - 1 \quad \text{(by (3.8))}$$

$$\leq \varlimsup_{\alpha \to \infty} \sum_{i=1}^\beta p_i \left[\int_{l<r} \frac{\mathbf{1}_{(b_i \in (l,r])}}{H(r) - H(l)} dQ(l,r) + \gamma_H(b_{i*}, b_i) - 1 \right] \quad \text{(by (E.1))}$$

$$= \int_{[\tau_o, \tau]} \left[\int \frac{\mathbf{1}_{(b \in (l,r])}}{H(r) - H(l)} dQ(l,r) + \lim_{a \uparrow b} \gamma_H(a,b) - 1 \right] dF(b) \quad \text{(by the BCT)}$$

$$= 0 \quad \text{(by (E.3))}.$$

Thus we reach a contradiction $0 < \delta \leq 0$. This concludes the proof of Theorem 3.1. □

4. Asymptotic Normality

If $F(\tau) = 0$, the GMLE $\hat{F}(\tau) = 0$ w.p.1. If $F(\tau) < 1$, $F(t)$ is not identifiable for $t > \tau$. Thus it suffices to estimate F_τ defined by $F_\tau(t) = \begin{cases} F(t) & \text{if } t \leq \tau \\ F(\tau) & \text{if } \tau < t < \infty \\ 1 & \text{if } t = \infty \end{cases}$, and assume that $F(\tau) > 0$. Here $F_\tau \in \Theta$ but may not be a cdf and Theorem 3.1 does not require F be a cdf.

There are two equivalent forms for equation (3.1): $H = \mathcal{B}_H(Q)$ and $H = \mathcal{R}_H(F)$, where

$$\mathcal{B}_H(Q)(x) = \int_{l<x<r} \frac{H(x) - H(l)}{H(r) - H(l)} dQ(l,r) + \int_{r \leq x} dQ(l,r) \quad \text{(RHS of (3.1))}, \tag{4.1}$$

$$\mathcal{R}_H(F)(x) = \int_{l \leq x < r} \left\{ \frac{H(x) - H(l)}{H(r) - H(l)} [F(r) - F(l)] - [F(x) - F(l)] \right\} dG^*(l,r) + F(x), \tag{4.2}$$

$$dG^*(l,r) = \begin{cases} \pi_2 dP(V \leq l) + \pi_0 dG_T(l) & \text{if } r = \infty, \\ \pi_2 dP(U \leq r) & \text{if } l = -\infty, \\ \pi_2 dG(l,r) & \text{if } -\infty < l < r < \infty, \end{cases}$$

$$dQ(l,r) = [F(r) - F(l)] dG^*(l,r) = [F_\tau(r) - F_\tau(l)] dG^*(l,r) \text{ if } l < r. \tag{4.3}$$

Lemma 4.1 . $\mathcal{B}_{H_n}(Q_n - Q) = \mathcal{R}_{H_n}(H_n - F_\tau)$ *for each SCE H_n which satisfies (2.3).*

The proof of the lemma is in Appendix B.

Let $\mathcal{D}$ be the collection of all real-valued functions h defined on $[-\infty, \infty]$ that are right-continuous, have left limits at each point and satisfy that

$$\forall \, a < b \leq \infty, \; F_\tau(a-) = F_\tau(b) \Rightarrow h(a-) = h(b). \tag{4.4}$$

Define $\mathcal{D}_0 = \{h \in \mathcal{D} : F(x) = 0 \Rightarrow h(x) = 0; F_\tau(x-) = 1 \Rightarrow h(x-) = 0\}$. Verify that $(\mathcal{D}, ||\cdot||)$ and $(\mathcal{D}_0, ||\cdot||)$ are both Banach spaces. Let $(\mathcal{D}_2, ||\cdot||)$ be a Banach space of real-valued functions defined on $[-\infty, \infty]^2$ such that the Banach space contains all bivariate cdfs, where $||g|| = \sup_{x,y} |g(x,y)|$. Note that AS1– AS3 are basically assumptions on (F, G, G_T). We say (H, G, G_T) satisfies AS1 etc., if $H \in \Theta$ and H replaces the role

7

of F in AS1 etc. Let $\Theta_o = \{H \in \Theta \cap \mathcal{D} : \mathcal{S}_H \subset \mathcal{S}_F, (H, G, G_T) \text{ satisfies AS1 or AS3}\}$. For each $H \in \Theta_o$, $\mathcal{R}_H(\cdot)$ and $\mathcal{B}_H(\cdot)$ are linear operators on $\mathcal{D}$ and $\mathcal{D}_2$, respectively.

Theorem 4.1. *Suppose that AS1, AS2 and (2.3) hold. Then $\mathcal{R}_{F_\tau}^{-1}$ exists as a bounded operator from $\mathcal{D}$ to $\mathcal{D}$ and the SCE satisfies*

$$\sqrt{n}(H_n - F_\tau) \overset{\mathcal{D}}{\longrightarrow} \mathcal{R}_{F_\tau}^{-1} \mathcal{B}_{F_\tau}(W) \text{ in } \mathcal{D}, \tag{4.5}$$

where W is the Gaussian process specified by $\sqrt{n}(Q_n(l, r) - Q(l, r)) \overset{\mathcal{D}}{\longrightarrow} W$.

We first state 3 more lemmas, with their proofs relegated to Appendix B.

For a $\tilde{F} \in \Theta_o$, let C_k be the collection of all the distinct points among $c_{k,i}$s, where $c_{k,i} = \inf\{x : \tilde{F}(x) \geq i/2^k\}$, $i = 0, ..., 2^k$, $k \geq 1$. Let F_k be a step function in Θ_o such that $F_k(c) = \tilde{F}(c)$ for each $c \in C_k$ and its discontinuity points belong to C_k. Denote $\mathcal{D}_k$ ($\mathcal{D}_{k0}$) the subclass of $\mathcal{D}$ ($\mathcal{D}_0$) such that each member is a step function with the collection of discontinuity points being a subset of $\mathcal{S}_{F_k}$. Obviously, D_k, C_k and F_k depend on $\tilde{F}$.

Lemma 4.2. *If $\tilde{F} \in \Theta_o$, then the linear operator $\mathcal{R}_{F_k}^{-1}$ exists as a map from $\mathcal{D}_k$ onto $\mathcal{D}_k$.*

Lemma 4.3. *Assume that AS1, AS2 and (2.3) hold. For each $\omega \in \Omega$, $H_n \in \Theta_o$.*

Lemma 4.4. *If $\tilde{F} \in \Theta_o$, then $\|\mathcal{R}_{F_k}^{-1}(\cdot)\| \leq 1$ for all possible k.*

Proof. We give the proof of asymptotic normality in 4 steps.

Step 1 (Existence of $\mathcal{R}_{\tilde{F}}^{-1}$, $\tilde{F} \in \Theta_o$, as a linear operator from $\mathcal{D}$ to $\mathcal{D}$). For each $g \in \mathcal{D}$ and $k \geq 1$, let $g_k \in \mathcal{D}_k$ be such that $g_k(x) = g(x)$ if $x \in C_k$. Then $\|g_k - g\| \to 0$, since $\mathcal{S}_{g_k}$ and $\mathcal{S}_g \subset \mathcal{S}_{\tilde{F}}$ and $C = \cup_k C_k$ is dense in $\mathcal{S}_{\tilde{F}}$. By Lemma 4.2, $\mathcal{R}_{F_k}^{-1}$ exists, so there exists a unique $h_k \in \mathcal{D}_k$ such that $g_k = \mathcal{R}_{F_k}(h_k)$. $\forall K > k$ and $\forall h \in \mathcal{D}$, $\mathcal{D}_k \subset \mathcal{D}_K$, $\|F_k - F_K\| \leq 1/2^k$ and $\mathcal{R}_{F_k}(h) - \mathcal{R}_{F_K}(h)$ converges to 0 as $k \to \infty$ by the BCT. $\|\mathcal{R}_{F_k}^{-1}(\cdot)\| \leq 1$ by Lemma 4.4, thus $\lim_{k \to \infty}[\mathcal{R}_{F_k}^{-1}(h) - \mathcal{R}_{F_K}^{-1}(h)] = 0 \ \forall \ h \in \mathcal{D}_k$ and $\forall \ k \geq 1$. Furthermore,

$$\|h_k - h_K\| \leq \|\mathcal{R}_{F_k}^{-1}(g_k) - \mathcal{R}_{F_K}^{-1}(g_k)\| + \|\mathcal{R}_{F_K}^{-1}(g_k) - \mathcal{R}_{F_K}^{-1}(g_K)\|$$

$$\leq \|\mathcal{R}_{F_k}^{-1}(g_k) - \mathcal{R}_{F_K}^{-1}(g_k)\| + \|\mathcal{R}_{F_K}^{-1}\| \cdot \|g_k - g_K\| \to 0 \text{ as } k \to \infty,$$

by the assumption $\|g_k - g\| \to 0$, Lemmas 4.2 and 4.4, and the BCT. That is, $\|h_k\|$ is a Cauchy sequence. Since $\mathcal{D}$ is a Banach space, there is a function $h_o \in \mathcal{D}$ such that $\|h_k - h_o\| \to 0$. By the BCT, $g = \lim_{k \to \infty} \mathcal{R}_{F_k}(h_k) = \mathcal{R}_{\tilde{F}}(h_o)$. Define $h_o = \mathcal{R}_{\tilde{F}}^{-1}(g)$.

Step 2 (Strong continuity of $\{\mathcal{R}_H^{-1} : H \in \Theta_o\}$). Let $g'_m \in \mathcal{D}$ and $H_m \in \Theta_o$ be such that $\|g'_m - g\| \to 0$ and $\|H_m - F_\tau\| \to 0$ as $m \to \infty$. Then

$$\|\mathcal{R}_{H_m}^{-1}(g'_m) - \mathcal{R}_{F_\tau}^{-1}(g)\| \leq \|\mathcal{R}_{H_m}^{-1}(g'_m) - \mathcal{R}_{F_\tau}^{-1}(g'_m)\| + \|\mathcal{R}_{F_\tau}^{-1}(g'_m) - \mathcal{R}_{F_\tau}^{-1}(g)\|$$

$$\leq \|\mathcal{R}_{H_m}^{-1} - \mathcal{R}_{F_\tau}^{-1}\| \cdot \|g'_m\| + \|\mathcal{R}_{F_\tau}^{-1}\| \cdot \|g'_m - g\| \to 0 \text{ as } m \to \infty.$$

Step 3 (Strong continuity of $\{\mathcal{B}_H : H \in \Theta_o\}$). Let h be a simple function in $\mathcal{D}_2$. It follows from (4.1) and the BCT that $\mathcal{B}_H(h) \to \mathcal{B}_{F_\tau}(h)$ in $\mathcal{D}$ as $H \to F_\tau$. Since $\|\mathcal{B}_H\| \leq 4 \ \forall \ H \in \Theta_o$ and the collection of simple functions is dense in $\mathcal{D}_2$, we have strong continuity.

Step 4 (Conclusion). By Lemma 4.3, $H_n \in \Theta_o$. Thus $\mathcal{R}_{H_n}^{-1}$ exists by Step 1. It follows that $\sqrt{n}(H_n - F_\tau)) = \mathcal{R}_{H_n}^{-1} \mathcal{B}_{H_n}(\sqrt{n}[Q_n - Q])$ by Lemma 4.1. By Theorem 2.1 $\lim_{n \to \infty} |H_n(x) - F_\tau(x)| = 0$ a.s. By Steps 2 and 3, $\{\mathcal{F}_H = \mathcal{R}_H^{-1} \mathcal{B}_H : H \in \Theta_o\}$ is strongly continuous. As a consequence of the above 4 statements, and the Banach-Steinhaus theorem, $\sup\{\|\mathcal{F}_{H_n}(h) - \mathcal{F}_{F_\tau}(h)\| : h \in A(\epsilon)\} \to 0$ a.s. as $n \to \infty$ and then $\epsilon \to 0+$ for all compact set $A \subset \mathcal{D}_2$, where $A(\epsilon) = \{h \in \mathcal{D}_2 : \|h - h'\| < \epsilon \text{ for some } h' \in A\}$. By the central limit theorem, $W_n = \sqrt{n}[Q_n - Q] \overset{\mathcal{D}}{\longrightarrow} W$ in $\mathcal{D}_2$, $\{W_n\}$ is uniformly tight (Pollard, 1984, p.81). As a consequence, $\|\sqrt{n}(H_n - F_\tau) - \mathcal{F}_{F_\tau}(W_n)\| = \|(\mathcal{F}_{H_n} - \mathcal{F}_{F_\tau})(W_n)\| = o_p(1)$, which implies (4.5) by the continuous mapping theorem (Pollard, 1984, p. 70). □

Remark 3. *Our proof of the normality (not the consistency) relies on the form (2.3). It can be shown that Theorem 4.1 are actually true without (2.3), and Theorem 2.1 (not Theorem 4.1) are true without (AS1.a). For the sake of simplicity, we skip the details.*

Theorem 2.2 is a consequence of Theorem 4.1.

Remark 4. *Under assumptions AS1 and AS2, H_n is also efficient. The proof is analogous to that of Theorem 3 of Gu and Zhang (1993) and is skipped here.*

Appendix A

We shall prove Lemmas 3.3 and 3.4. A lemma is needed to prove Lemma 3.3.

Lemma A.1. *Assume that AS1 or AS3 holds. Let $\psi(H)$ be a limit of $\{\psi_\alpha(H)\}$, $H \in \Theta$. Then $\psi(H) = 0$ if and only if (1) $H(t) = F(t)$ and $H(t-) = F(t-) \; \forall \; t \in \mathcal{S}_F \cap \cup_\alpha B_\alpha \cap (-\infty, \tau)$, (2) $H(\tau-) = F(\tau-)$ if $F(\tau-) < 1$ and (3) $H(\tau) = F(\tau)$ if $F(\tau-) < 1$ and AS1 holds. Moreover, $\psi(H) \leq 0$.*

Proof. ($\Rightarrow$) Verify that $\psi_\alpha(F) = 0$ for all α by AS1.a, and thus $\lim_{\alpha \to \infty} \psi_\alpha(F) = 0$. Then conditions (1) – (3) above imply that $\psi_\alpha(H) = \psi_\alpha(F) = 0$ for all $\alpha \geq 1$. Thus $\psi(H) = 0$.

($\Leftarrow$) We first show that $\psi(H) = 0$ implies condition (1). It suffices to show that $\psi(H) < 0$ if for some $t_0 \in \mathcal{S}_F \cap \cup_\alpha B_\alpha \cap (-\infty, \tau)$ either (1.a) $H(t_0) \neq F(t_0)$ or (1.b) $H(t_0-) \neq F(t_0-)$. Condition (1.a) implies that for each sufficient large α, there is a point $b_h \in \mathcal{S}_F \cap B_\alpha$ such that $b_h = t_0$. Verify that

$$
\begin{aligned}
\psi_\alpha(H) &= E\{E(\ln(\mu_H(I_\alpha)/\mu_F(I_\alpha))|U_\alpha, V_\alpha, T_\alpha, \mathcal{K})\} \\
&= \pi_2 \int \int f_{\alpha,2}(z,y) dG_\alpha(z,y) + \pi_0 \int f_{\alpha,0}(t) dG_{T,\alpha}(t), \quad \text{where}
\end{aligned}
\tag{A.1}
$$

$$
f_{\alpha,2}(z,y) = F(z)\ln\frac{H(z)}{F(z)} + [F(y) - F(z)]\ln\frac{H(y) - H(z)}{F(y) - F(z)} + [1 - F(y)]\ln\frac{1 - H(y)}{1 - F(y)},
$$

$$
f_{\alpha,0}(b_j) = F(b_1)\ln\frac{H(b_1)}{F(b_1)} + \sum_{k>1}^{j} \mu_F(\lfloor b_{k*}, b_k]) \ln\frac{\mu_H(\lfloor b_{k*}, b_k])}{\mu_F(\lfloor b_{k*}, b_k])} + [1 - F(b_j)]\ln\frac{1 - H(b_j)}{1 - F(b_j)},
$$

and b_k, t_0 and $b_j \in B_\alpha$. Note t_0 is fixed but the index h of $b_h = t_0$ depends on α. Define

$$
g(k,t) = \begin{cases} F(t_0)\ln\frac{H(t_0)}{F(t_0)} + [1 - F(t_0)]\ln\frac{1-H(t_0)}{1-F(t_0)} & \text{if } t_0 \leq t, \\ 0 & \text{otherwise.} \end{cases}
\tag{A.2}
$$

Then $0 = \ln\left[F(t_0)\frac{H(t_0)}{F(t_0)} + [1 - F(t_0)]\frac{1-H(t_0)}{1-F(t_0)}\right] > F(t_0)\ln\frac{H(t_0)}{F(t_0)} + [1 - F(t_0)]\ln\frac{1-H(t_0)}{1-F(t_0)} = g(0,t)$, for $t \geq t_0$, as $-\ln(\cdot)$ is strictly convex and $F(t_0) \neq H(t_0)$. Moreover, $P\{T \text{ or } V \geq t_0\} > 0$ as $\pi_0 > 0$ and $t_0 \in (-\infty, \tau)$. It follows from the above two statements that

$$
P\{0 > g(\mathcal{K}, T)\} > 0.
\tag{A.3}
$$

It is obvious that (1.a.1) $g(2,t) \geq f_{\alpha,2}(u,v)$ for each (u,v,t) and (1.a.2) $0 = g(0,t) \geq f_{\alpha,0}(t)$ for $t < t_0$. We shall show that, (1.a.3) $g(0,t) > f_{\alpha,0}(t)$, for $t = b_k \geq t_0$, where $b_k \in B_\alpha$ and α is sufficiently large. Let $\int g dG_\alpha^w = \pi_2 \int \int g(2,t) dG_\alpha(u,v) + \pi_0 \int g(0,t) dG_{T_\alpha}(t)$, and define $\int g dG^w$ in an obvious way. Then (1.a.1), (1.a.2) and (1.a.3) imply that $\int g dG_\alpha^w \geq \psi_\alpha(H)$. Since dG_α^w converges to dG^w setwisely by observing that dG_α (dG_{T_α}) converges to dG (dG_T) setwisely and $g(k,t)$ is a binary function in (u,v,t,k), the desired result follows from (A.3) and $0 > \int g dG^w = \varlimsup_{\alpha \to \infty} \int g dG_\alpha^w \geq \varlimsup_{\alpha \to \infty} \psi_\alpha(H) \geq \psi(H)$.

We now establish (1.a.3). Let $t_o = b_h \leq b_j = t$ for some integer α_o. It is easy to see by our construction that $B_{\alpha_1} \subset B_{\alpha_2}$ if $\alpha_1 < \alpha_2$ and hence $t_0, t \in B_\alpha$ for all $\alpha \geq \alpha_o$. For each $z = b_i \in B_{\alpha_1}$ such that $z < t_0$, verify

(i)
$$
\begin{aligned}
g(0,t) = F(t_0)\ln\Big\{&\frac{H(z)}{F(z)}\frac{F(z)}{F(t_0)} + \frac{(H(t_0) - H(z))}{(F(t_0) - F(z))}\frac{F(t_0) - F(z)}{F(t_0)}\Big\} \\
&+ [1 - F(t_0)]\ln\Big\{\frac{(H(t) - H(t_0))}{(F(t) - F(t_0))}\frac{F(t) - F(t_0)}{1 - F(t_0)} + \frac{1 - H(t)}{1 - F(t)}\frac{1 - F(t)}{1 - F(t_0)}\Big\},
\end{aligned}
$$

(ii)
$$
\ln\frac{H(b) - H(a)}{F(b) - F(a)} \geq \frac{F(x) - F(a)}{F(b) - F(a)}\ln\frac{H(x) - H(a)}{F(x) - F(a)} + \frac{F(b) - F(x)}{F(b) - F(a)}\ln\frac{H(b) - H(x)}{F(b) - F(x)}
$$
for all $x \in (a,b)$.

In view of (i) and (ii), (1.a.3) follows by an induction argument.

9

Now consider condition (1.b). If t_0 is a point satisfying condition (1.b), then either (1.b.1) $t_0 \in \mathcal{S}_F \cap (\cup_\alpha B_\alpha) \cap (\cup_\alpha B_\alpha^*)$, where $B_\alpha^* = \{x : x = b_{i*} > b_{i-1}, b_i, b_{i-1} \in B_\alpha\}$, or (1.b.2) $t_0 \in \mathcal{S}_F \cap (\cup_\alpha B_\alpha) \cap (\cup_\alpha B_\alpha^*)^c$, where A^c is the complement of the set A.

First assume (1.b.1). For each sufficiently large α, there exists a $b_{h*} = t_0 \in B_\alpha^*$. Thus replacing t_0 by t_0- in the proof for situation (1.a) yields $\psi(H) < 0$.

On the other hand, in view of (3.3), (1.b.2) implies that $F(t_0-) > F(t)$ for each $t < t_0$ and hence there exists a sequence of points $x_i \in \mathcal{S}_F \cap \cup_\alpha (B_\alpha \cup B_\alpha^*)$ such that $x_i \uparrow t_0$ with either $H(x_i) \neq F(x_i)$ (if $x_i = b_{j*} = b_{j-1}$) or $H(x_i-) \neq F(x_i-)$ (if $x_i = b_{j*} > b_{j-1}$). In either case, it reduces to situation (1.a) or (1.b.1). Thus, we have $\psi(H) < 0$. This concludes the proof for condition (1.b).

The proofs for conditions (3) and (2) are similar to that for conditions (1.a) and (1.b), respectively, except in the proof for condition (3) replacing in the above proof the statement $P\{t_0 \leq T\} > 0$ by $P\{T \text{ or } V = \tau_t\} > 0$ (as AS1 holds). We omit the details.

Verify that we actually show that either $\psi(H) = 0$ or $\psi(H) < 0$. Thus $\psi(H) \leq 0$. $\square$

Proof of Lemma 3.3. Statement (2) follows from the last statement in Lemma A.1. To prove statement (1), it suffices to show that conditions (1), (2) and (3) in Lemma A.1 imply $H(x) = F(x) \, \forall \, x \leq \tau$, i.e. the sufficient and necessary condition in Lemma 3.3.

If x is a discontinuity point of F and $x \leq \tau$, then there exists an integer N such that $F(x) - F(x-) > 2^{-\alpha}$ for all $\alpha \geq N$. This implies that x is a certain $j2^{-N} \times 100$ percentile of μ and thus $x \in B_\alpha \cap \mathcal{S}_F$. It follows that $\mathcal{S}_F \cap \cup_\alpha B_\alpha$ contains all discontinuity points of F which belong to $(-\infty, \tau]$. Thus conditions (1), (2) and (3) of Lemma A.1 imply $H(x) = F(x)$.

Suppose now x is a continuity point of F. Let $u_x = \inf\{y : F(y) = F(x)\}$ and $v_x = \sup\{y : F(y) = F(x)\}$. If both u_x and v_x belong to $\mathcal{S}_F \cap \cup_\alpha B_\alpha$, we are done, as $F(x) = F(u_x) = H(u) \leq H(x) \leq H(v_x-) = F(v_x-) = F(x)$ by conditions (1), (2) and (3) in Lemma A.1.

If neither u_x nor v_x belongs to $\mathcal{S}_F \cap \cup_\alpha B_\alpha$, then from the above discussion both u_x and v_x are continuous support points of F satisfying $F(u_x) = F(v_x) = F(x)$, and there exist two sequences of support points of F, say $\{x_i\}_{i \geq 1}$ and $\{y_j\}_{j \geq 1}$, which are contained in $\mathcal{S}_F \cap \cup_\alpha B_\alpha$ such that $x_i \uparrow u_x$ and $y_j \downarrow v_x$. Consequently, $F(x_i) = H(x_i) \leq H(x) \leq H(y_j) = F(y_j)$ by conditions (1), (2) and (3) in Lemma A.1. This yields $H(x) = F(x)$ as $F(x_i) \to F(u_x)$ and $F(y_j) \to F(v_x)$.

For simplicity, we skip the proof for the case that only u_x or v_x belongs to $\mathcal{S}_F \cap (\cup_\alpha B_\alpha)$. This concludes the proof of the lemma. $\square$

A lemma is needed for proving Lemma 3.4.

Lemma A.2. *Suppose that H is a solution of (3.1) and A is an interval $(a, b] \subset (-\infty, \tau]$. Then $\mu_F(A) > 0 \Rightarrow \mu_H(A) > 0$.*

Proof. Equation (3.1) is equivalent to

$$\mu_H((a, b]) = \int_{l < r} \frac{\mu_H((a, b] \cap (l, r])}{H(r) - H(l)} dQ(l, r) + P(X \in (a, b], X \leq T, \mathcal{K} = 0). \qquad (A.4)$$

If H is a solution to (3.1), then for each interval $A = (a, b] \subset (-\infty, \tau]$ such that $\mu_F(A) > 0$, we have $\mu_H(A) \geq P(X \in A, X \leq T, \mathcal{K} = 0) > 0$ by the assumption $\pi_0 > 0$ and $b \leq \tau$. This concludes the proof of the lemma. $\square$

Proof of Lemma 3.4. Assume that H is a solution of (3.1) and $b \in \mathcal{L}_F$. By Lemma A.2, $\mu_H((a, b]) > 0 \, \forall \, a < b$. Dividing both sides of equation (A.4) by $\mu_H((a, b])$ yields

$$1 = \int_{l < r} \frac{\mu_H((a, b] \cap (l, r])}{[H(b) - H(a)][H(r) - H(l)]} dQ(l, r) + \frac{P(X \in (a, b], X \leq T, \mathcal{K} = 0)}{H(b) - H(a)}, \quad a < b. \qquad (A.5)$$

For each $a < b$, (A.5) yields (E.2) as the two summands in (A.5) are nonnegative.

Denote $\partial H(a, b, l, r) = \begin{cases} 0 & \text{if } \mu_H((a, b] \cap (l, r]) = 0, \\ \frac{\mu_H((a, b] \cap (l, r])}{[H(b) - H(a)][H(r) - H(l)]} & \text{otherwise.} \end{cases}$

For each pair (l, r) such that $l < b \leq r$, we have $H(r) - H(l) > 0$ by Lemma A.2. Moreover,

$$\partial H(a, b, l, r) \uparrow \frac{\mathbf{1}_{(b \in (l, r])}}{H(r) - H(l)} \text{ as } a \uparrow b \text{ if } b \in (l, r), \text{ and } \partial H(a, b, l, r) \downarrow 0 \text{ as } a \uparrow b \text{ if } b > r.$$

10

Thus by the monotone convergence theorem, as $a \uparrow b$, we have

$$\int \partial H(a,b,l,r)dQ = \int_{b \in (l,r]} \partial H(a,b,l,r)dQ + \int_{b>r} \partial H(a,b,l,r)dQ \to \int \frac{1_{(b \in (l,r])}}{H(r)-H(l)}dQ.$$

The desired equation (E.3) follows from (A.5), (E.2) and the above equation.

Assume now $0 < F(\tau) < 1$ and AS1 holds. By AS1.a, $l \leq \tau_t < r \Rightarrow r = \infty$, thus

$$\mu_H((\tau,\infty]) = \int_{l \leq \tau < r} \frac{H(r)-H(\tau)}{H(r)-H(l)}dQ(l,r) \qquad \text{(by AS1 and (A.4))} \qquad (A.6)$$

$$\geq \int_{l=\tau<r} dQ(l,r) = (1-F(\tau))P(L=\tau) > 0 \quad \text{(by AS1.b)}.$$

Dividing both sides of equation (A.6) by $\mu_H((\tau_t,\infty])$ yields (E.1) under AS1.

On the other hand, assume that $0 < F(\tau) < 1$ and AS3 holds. Note that even if we encounter $\frac{0}{0}$, $\frac{\mu_H((x_o,\infty]\cap(l,r])}{H(r)-H(l)}1_{[x_o \leq l < r)} = 1$ by convention. By AS3, $P(x < T < \tau_t) > 0$ for each $x < \tau_t$. (A.4) yields

$$\mu_H((x,\infty]) = \int_{l<r} \frac{\mu_H((x,\infty] \cap (l,r])}{H(r)-H(l)}dQ(l,r) \qquad (A.7)$$

$$\geq (1-F(\tau_t))P(T \in (x,\tau_t]) > 0, \quad \forall\, x \in [x_o,\tau_t).$$

Then dividing both sides of (A.7) by $\mu_H((x,\infty])$ and taking limits yield

$$1 = \lim_{x \uparrow \tau_t} \int_{l<r} \frac{\mu_H((x,\infty]\cap(l,r])}{\mu_H((x,\infty))(H(r)-H(l))}dQ(l,r) = \int \frac{1_{(l \leq \tau_t < r)}}{H(r)-H(l)}dQ(l,r)$$

(as $\tau_v \leq \tau_t$ and thus $l \leq \tau_t < r \Rightarrow r = \infty$), which is (E.1). $\square$

Appendix B

In this appendix, we prove lemmas in Section 4.

Proof of Lemma 4.1. Theorem 3.1, (4.1), (4.2) and (2.3) yield $\mathcal{B}_{H_n}(Q_n)(x) = H_n(x)$ and $\mathcal{R}_{F_\tau}(F_\tau)(x) = F_\tau(x)\ \forall\, x$. Furthermore,

$$\int_{l \leq x < r} \frac{H_n(x)-H_n(l)}{H_n(r)-H_n(l)}\{[H_n(r)-F_\tau(r)] - [H_n(l)-F_\tau(l)]\}dG^*(l,r)$$

$$= \int_{l \leq x < r}[H_n(x)-H_n(l)]dG^*(l,r) - \int_{l \leq x < r} \frac{H_n(x)-H_n(l)}{H_n(r)-H_n(l)}[F_\tau(r)-F_\tau(l)]dG^*(l,r)$$

$$= \int_{l \leq x < r}[H_n(x)-H_n(l)]dG^*(l,r) - \int_{l \leq x < r} \frac{H_n(x)-H_n(l)}{H_n(r)-H_n(l)}dQ(l,r) \quad \text{(by (4.3))}$$

$$= \int_{l \leq x < r}[H_n(x)-H_n(l)]dG^*(l,r) - \mathcal{B}_{H_n}(Q)(x) + P\{R \leq x\} \quad \text{(by (4.1))}$$

$$= \int_{l \leq x < r}[H_n(x)-H_n(l)]dG^*(l,r)$$
$$\quad - \mathcal{B}_{H_n}(Q)(x) + [\mathcal{B}_{H_n}(Q_n)(x)-H_n(x)] \qquad \text{(since } \mathcal{B}_{H_n}(Q_n) = H_n)$$
$$\quad + [F_\tau(x) - \int_{l \leq x < r} F_\tau(x)-F_\tau(l)dG^*(l,r)] \qquad (= P\{R \leq x\} \text{ as } F_\tau = \mathcal{R}_{F_\tau}(F_\tau))$$

$$= \mathcal{B}_{H_n}(Q_n - Q)(x) - [H_n(x)-F_\tau(x)] + \int_{l \leq x < r}\{[H_n(x)-F_\tau(x)] - [H_n(l)-F_\tau(l)]\}dG^*.$$

Translating certain terms in the first and last expressions of the above equations yields

$$\mathcal{B}_{H_n}(Q_n - Q)(x)$$

11

$$= \int_{l \leq x < r} \frac{H_n(x) - H_n(l)}{H_n(r) - H_n(l)} \{[H_n(r) - F_\tau(r)] - [H_n(l) - F_\tau(l)]\} dG^*(l, r)$$

$$+ [H_n(x) - F_\tau(x)] - \int_{l \leq x < r} \{[H_n(x) - F_\tau(x)] - [H_n(l) - F_\tau(l)]\} dG^*(l, r)$$

$$= \int_{l \leq x < r} \left(\frac{H_n(x) - H_n(l)}{H_n(r) - H_n(l)} \{[H_n(r) - H_n(l)] - [F_\tau(r) - F_\tau(l)]\} \right.$$

$$\left. - \{[H_n(x) - H_n(l)] - [F_\tau(x) - F_\tau(l)]\} \right) dG^*(l, r) + [H_n(x) - F_\tau(x)]$$

$$= \mathcal{R}_{H_n}(H_n - F_\tau)(x) \qquad \text{(by (4.2))}. \ \square$$

Lemma B.1. *If $\tilde{F} \in \Theta_o$ and $\mathcal{R}_{\tilde{F}}(h) = 0$, where $h \in \mathcal{D}$, then $h \in \mathcal{D}_0$.*
Proof. For each $h \in \mathcal{D}$, by (4.2),

$$\mathcal{R}_{\tilde{F}}(h)(x) = \int_{l \leq x < r} \left\{ \frac{\tilde{F}(x) - \tilde{F}(l)}{\tilde{F}(r) - \tilde{F}(l)} [h(r) - h(l)] - [h(x) - h(l)] \right\} dG^*(l, r) + h(x). \qquad (B.1)$$

If $F(x) = 0$, then $h = h(x)$ on $(-\infty, x]$ by (4.4). Thus
$0 = \mathcal{R}_{\tilde{F}}(h)(x) = -\int_{l \leq x < r}[h(x) - h(l)]dG^*(l, r) + h(x) = h(x)$.
Moreover, if $F_\tau(x-) = 1$, then $h = h(x-)$ on $[x, \infty]$ by (4.4), and
$0 = \mathcal{R}_{\tilde{F}}(h)(x-) = \int_{l < x \leq r}[h(r) - h(x-)]dG^*(l, r) + h(x-) = h(x-)$. Thus $h \in \mathcal{D}_0$. $\square$

Proof of Lemma 4.2. Note $\tilde{F} \in \Theta_o$. Since $\mathcal{R}_{F_k}$ is a linear operator on the finite dimensional linear space $\mathcal{D}_k$, it suffices to show (1) $\mathcal{R}_{F_k}$ is 1-1 and (2) $\mathcal{R}_{F_k}(\mathcal{D}_k) \subset \mathcal{D}_k$.

Step (1). Suppose $\mathcal{R}_{F_k}(h) = 0$, where $h \in \mathcal{D}_k$. We shall show that $h = 0$. Denote $\alpha = \sum_{c \in C_k} |h(c) - h(c-)|$ and $m = \min\{m_c : m_c = F_k(c) - F_k(c-) > 0, c \in C_k\}$. Note that $m > 0$ and α is finite as C_k contains finitely many points. Choose $\gamma > 0$ such that $\gamma\alpha < m$. Let $H = F_k + \gamma h$. Since $\mathcal{R}_{F_k}(h) = 0$ and $\tilde{F} \in \Theta_o$, $h \in \mathcal{D}_{k0}$ by Lemma B.1. As consequences, (1) $H(\tau_o -) = F_k(\tau_o -) + 0 = 0$ and $H(\infty) = F_k(\infty) + 0 = 1$; (2) $H(c) > H(c-)$ for all $c \in C_k$ [as $H(c) - H(c-) > m - \gamma(h(c) - h(c-)) \geq m - \gamma\alpha > 0$]; (3) $H(x) \in \mathcal{D}_k$.

It follows from statements (1), (2) and (3) that $H = F_k + \gamma h \in \Theta_o \cap \mathcal{D}_k$. Then $\mathcal{R}_{F_k}(H)(x) = \mathcal{R}_{F_k}(F_k)(x) + \mathcal{R}_{F_k}(\gamma h)(x) = F_k(x) + 0$ for each x. That is $F_k = \mathcal{R}_{F_k}(H)$. Note that (H, G, G_T) satisfies AS1 or AS3 as $H \in \Theta_o$. Thus $F_k = H = F_k + \gamma h$ by Theorem 3.1, which implies $h = 0$ as $\gamma > 0$. As a consequence, $\mathcal{R}_{F_k}(\cdot)$ is 1-1.

Step (2). It suffices to show that $\mathcal{A} = \mathcal{R}_{F_k}(h)(b) - \mathcal{R}_{F_k}(h)(a) = 0$ if $h \in \mathcal{D}_k$ and $\mu_{F_k}((a, b]) = 0$. Define $\mu_h((a, b]) = h(b) - h(a)$. By definition of $\mathcal{D}_k$, $\mu_h((a, b]) = 0$. Then

$$\mathcal{A} = \int_{l < r} \left[\frac{\mu_{F_k}((l, r] \cap (a, b])}{\mu_{F_k}((l, r])} \mu_h((l, r]) - \mu_h((l, r] \cap (a, b]) \right] dG^*(l, r) + \mu_h((a, b]) = 0. \ \square$$

Proof of Lemma 4.3. We fix $\omega \in \Omega$, as H_n is random. We shall verify that H_n satisfies the properties of Θ_o. First, by AS2 $\tau_v \leq \tau_t$.

If $a < b$ and $F(a-) = F(b)$, then $[a, b] \cap \mathcal{S}_F = \emptyset$. It follows that $\{R_1, ..., R_n\} \cap [a, b] = \emptyset$ by AS2 and thus H_n satisfies (4.4) and $H_n \in \mathcal{D}$ by Convention (2.3).

If $H_n(\tau_t -) = 1$ then (H_n, G, G_T) trivially satisfies AS1 and thus $H_n \in \Theta_o$. Moreover, if $F(\tau_t -) < 1$, then P(T or $V = \tau_t$) > 0 and thus (H_n, G, G_T) also satisfie AS1. It follows that $H_n \in \Theta_o$. Hence, WLOG, we can assume that $F(\tau_t -) = 1$. and $H_n(\tau_t -) < 1$. Then either $P(R = \tau_t) > 0$ or $P(R = \tau_t) = 0$ If $P(R = \tau_t) > 0$, then P($V = \tau_t$) > 0 as $F(\tau_t -) = 1$. I.e., (H_n, G, G_T) satisfies AS1 and $H_n \in \Theta_o$. If $P(R = \tau_t) = 0$ then with probability one $R_i \neq \tau_t$. WLOG, we can assume that $R_i \neq \tau_t$. Let x_o be the largest R_i that is smaller than τ_t. Then (2.3) implies that $\mu_{H_n}([x_o, \tau_t]) = 0$. Moreover, $\tau_v \leq \tau_t$ by AS1. Hence, (H_n, G, G_T) satisfies AS3. It follows that $H_n \in \Theta_o$. $\square$

Proof of Lemma 4.4. Let $o_1, ..., o_m$ be all the discontinuity points of F_k. Then $\mathcal{D}_k$ is an m-dimensional linear space. Define $h_i(x) = 1_{(x \geq o_i)}$. We shall show that

$$\mathcal{A}_{F_k} = \mathcal{R}_{F_k}(h_i)(o_j) - \mathcal{R}_{F_k}(h_i)(o_j -) \geq 0 \text{ for each } j \text{ and for each } h_i. \qquad (B.2)$$

12

Verify that $\mathcal{R}_{F_k}(h_i) \in \mathcal{D}_k$ (by Lemma 4.2), $\mathcal{R}_{F_k}(h_i)(0-) = 0$ and $\mathcal{R}_{F_k}(h_i)(\infty) = 1$. Then

$$\mathcal{R}_{F_k}(h_i), \ i = 1, ..., m, \ \text{are a base of } \mathcal{D}_k \text{ and } \|\mathcal{R}_{F_k}(h_i)\| = 1, \qquad (B.3)$$

by Lemma 4.2, as $\mathcal{D}_k$ is an m-dimensional linear space, and

$$h_i, \ i = 1, ..., m, \ \text{are a base of } \mathcal{D}_k \text{ and } \|h_i\| = 1. \qquad (B.4)$$

(B.3) and (B.4) imply that $\|\mathcal{R}_{F_k}^{-1}\| = 1$.

The proof of the lemma will be completed after we prove (B.2). Letting $x = o_j$, $h = h_i$, $F = F_k$, (B.1) yields

$$\mathcal{R}_F(h)(x) = \int_{l \leq x < r} [\frac{F(x) - F(l)}{F(r) - F(l)} h(r) + \frac{F(r) - F(x)}{F(r) - F(l)} h(l)] dG^*(l, r) + \beta(x),$$

where $\beta(x) = \pi_0(1 - G_T(x))h(x)$. Moreover, $\{l \leq x- < r\} = \{l < x \leq r\}$ and

$$\begin{aligned}
\mathcal{A}_F &= \int_{l \leq x < r} [\frac{F(x) - F(l)}{F(r) - F(l)} h(r) + \frac{F(r) - F(x)}{F(r) - F(l)} h(l)] dG^*(l, r) \\
&\quad - \int_{l < x \leq r} [\frac{F(x-) - F(l)}{F(r) - F(l)} h(r) + \frac{F(r) - F(x-)}{F(r) - F(l)} h(l)] dG^*(l, r) + \beta(x) - \beta(x-) \\
&= \int_{l < x < r} \frac{(F(x) - F(x-))(h(r) - h(l))}{F(r) - F(l)} dG^*(l, r) + \int_{l=x<r} h(x) dG^*(l, r) \qquad (B.5) \\
&\quad - \int_{l<x=r} [\frac{F(x-) - F(l)}{F(x) - F(l)} h(x) + \frac{F(x) - F(x-)}{F(x) - F(l)} h(l)] dG^*(l, r) + \beta(x) - \beta(x-).
\end{aligned}$$

Replacing F and h by F_k and $\mathbf{1}_{(x \geq o_i)}$, respectively, equation (B.5) yields

$$\begin{aligned}
\mathcal{A}_{F_k} &\geq \int_{l=x<r} h(x) dG^*(l, r) - \int_{l<x=r} h(x) dG^*(l, r) + \beta(x) - \beta(x-) \\
&= \pi_0 P(T = x)h(x) + \pi_0(1 - G_T(x))h(x) - \pi_0(1 - G_T(x-))h(x-) \\
&= \pi_0(1 - G_T(x-))(h(x) - h(x-)) \\
&\geq 0, \ \text{which is (B.2).} \ \square
\end{aligned}$$

13

References

* Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T. and Silverman, E. (1955). An empirical distribution function for sampling incomplete information. *Ann. Math. Statist.,* 26, 641-647.

* Chang, M.N. and Yang, G. (1987). Strong consistency of a self-consistent estimator of the survival function with doubly-censored data. *Ann. Statist.,* 15, 1536-1547.

* Finkelstein, D.M. and Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics,* 41, 933-945.

* Gentleman, R. and Geyer, C. J (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika,* 81, 618-623.

* Groeneboom, P. and Wellner, J.A. (1992). Information bounds and nonparametric maximum likelihood estimation. *Birkhäuser Verlag, Basel.*

* Groeneboom, P. (1996). Inverse problem in statistics. Proceedings of the St. Flour Summer School in Probability, 1994. Lecture notes in Math. 1648, Springer Verlag, Berlin.

* Gu, M.G. and Zhang, C.H. (1993). Asymptotic properties of self-consistent estimator based on doubly censored data. *Ann. Statist.,* 21, 611-624.

* Huang, J. (1996). Efficient estimation for proportional hazards models with interval censoring. *Ann. Statist.,* 24, 540-568.

* Huang, J. (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica,* 9, 501-520.

* Li, L.X., Watkins, T. and Yu, Q.Q. (1997). An EM algorithm for smoothing the self-consistent estimator of survival functions with interval - censored data. *Scand. J. Statist.,* 24, 531-542.

* Odell, P.M., Anderson, K.M. and D'Agostino, R.B. (1992). Maximum likelihood estimation for interval-censored data using a Weibull-based accelerated failure time model. *Biometrics,* 48, 951-959.

* Petroni, G. R. and Wolfe, R. A. (1994). A two-sample test for stochastic ordering with interval-censored data. *Biometrics,* 50, 77-87.

* Peto, R. (1973). Experimental survival curves for interval-censored data. *Applied Statistics,* 22, 86-91.

* Pollard, D. (1984). Convergence of stochastic processes. *Springer-Verlag.* New York.

* Schick, A. and Yu, Q.Q. (1999). Consistency of the GMLE with mixed case interval-censored data. *Scand. J. Statist.,* 26.

* Turnbull, B.W. (1976). The empirical distribution function with arbitrary grouped, censored and truncated data. *JRSS,* B. 38, 290-295.

* Van De Geer, S. (1996). Rates of convergence for the maximum likelihood estimator in mixture models. *Nonpar. Statist.,* 293-310.

* Wellner, J.A. (1995). Interval censoring case 2: alternative hypotheses. In *Analysis of censored data,* Proceedings of the workshop on analysis of censored data, December 28, 1994-January 1, 1995, University of Pune, Pune, India. *IMS Lecture Notes, Monograph Series.*

* Wellner, J.A. and Zhan, Y. (1997). A hybrid algorithm for computation of the NPMLE from censored data. *JASA,* **92**, 945-959.

* Yu, Q.Q. and Li, L.X. (1994). On the strong consistency of the product limit estimator. *Sankhyā,* A, 56, 416-430.

* Yu, Q.Q., Li, L.X. and Wong, G.Y.C. (1998a). Asymptotic variance of the GMLE of a survival function with interval-censored data. *Sankhyā,* A. 60, 184-197.

* Yu, Q.Q., Schick, A., Li, L.X. and Wong, G.Y.C. (1998b). Asymptotic properties of the GMLE in the case 1 interval censorship model with discrete inspection times. *Canadian J. Statist.,* 26, 619-627.

* Yu, Q.Q., Schick, A., Li, L.X. and Wong, G.Y.C. (1998c). Asymptotic properties of the GMLE of a survival function with case 2 interval-censored data. *Statist. & Prob. Lett.,* 37, 223-228.

A MODIFIED GMLE WITH DOUBLY-CENSORED DATA

Qiqing Yu[1] and George Y. C. Wong [1]

Department of Mathematical Sciences, SUNY at Binghamton, NY 13902, USA

and

Strang Cancer Prevention Center, 428 E 72nd Street, NY 10021, USA

Abstract

We consider efficient estimation of a distribution function F of a random variable X with doubly-censored data. The double censorship model assumes that X and the random vector (Z, Y) are independent and $Z < Y$ with probability one, and that X is uncensored if $Z < X \leq Y$, right censored if $Y < X$ and left censored if $X \leq Z$. Let $K(x) = P(Z \leq x < Y)$ and let $\mathcal{B} = \{x : K(x-) = 0, F(x) > 0 \text{ and } F(x-) < 1\}$. Under the assumption $P(X \in \mathcal{B}) = 0$, we present an example that the generalized maximum likelihood estimator (GMLE) of F with doubly-censored data is not asymptotically normally distributed and is not asymptotically efficient, and we propose a modified GMLE. We conjecture that it is asymptotically normally distributed and asymptotically efficient under the assumption $P(X \in \mathcal{B}) = 0$. We give a proof under an additional assumption.

1

1. Introduction

We consider efficient estimation of a survival function with doubly-censored data. Let $X_1, X_2, \ldots, X_n$ be i.i.d. copies from a random survival time X, with a distribution function F. Let $(Z_1, Y_1), (Z_2, Y_2), \ldots, (Z_n, Y_n)$ be i.i.d. copies from a random vector (Z, Y), where $Z < Y$ with probability one. Assume that X and (Z, Y) are independent. For each i, $1 \le i \le n$, X_i is either observed if $Z_i < X_i \le Y_i$, or right censored if $X_i > Y_i$, or left censored if $X_i \le Z_i$. Thus the observation can be represented by a random interval $\mathcal{I}$, where

$$\mathcal{I} = \begin{cases} [X, X] & \text{if } Z < X \le Y, \\ (Y, \infty) & \text{if } Y < X, \\ (-\infty, Z] & \text{if } X \le Z. \end{cases} \tag{1.1}$$

This censoring scheme is called a double censorship model (DC model).

Doubly-censored data often arise in biomedical studies, reliability research, and many other fields. Examples of doubly-censored data can be found in Leiderman *et al* (1973), Samuelsen (1989) and Kim, De Gruttola and Lagakos (1993).

Turnbull (1974) proposes the generalized maximum likelihood estimator (GMLE) of F with doubly-censored data, and shows that the GMLE is a self-consistent estimator (SCE). Turnbull (1974), Chang and Yang (1987), Chang (1990) and Gu and Zhang (1993) show that the SCEs are consistent, asymptotically normally distributed and asymptotically efficient under certain regularity conditions. Denote

$K(x) = \mathrm{P}(Z \le x < Y),$

$P_c(x) = P\{X \text{ is not censored} | X = x\},$

$Q = \{x : F(x) > 0 \text{ and } F(x-) < 1\},$

$\mathcal{B} = \{x : K(x-) = 0, \ x \in Q\}.$

Turnbull (1974) assumes all random variables takes on finitely many values. Gu and Zhang (1993) make weaker assumptions, with a key assumption

$$K(x-) = P_c(x) > 0 \text{ for all } x \in Q. \tag{1.2}$$

Let Ω be the sample space and let $\mathcal{O}_F = \{x : x = X(\omega) \text{ for some } \omega \in \Omega\}$. Assumption (1.2) implies that

$$\mathcal{O}_F \supset Q \text{ and } K(x-) = P_c(x) \text{ for all } x \in Q, \tag{1.3}$$

2

which is not true for a discrete random variable X. The condition really needed in the proofs of Gu and Zhang (1993) is

$$K(x-) > 0 \text{ for all } x \in Q, \tag{1.4}$$

rather than (1.2). (1.4) is weaker than (1.2), as it does not imply (1.3).

A sufficient condition for F to be identifiable on the whole real line is $P(X \in \mathcal{B}) = 0$, since all SCEs are consistent under this assumption (Yu and Li (1998)). It is easy to see $P(X \in \mathcal{B}) = 0$ is weaker than (1.4) since (1.4) implies that $\mathcal{B}$ is empty, thus $P(X \in \mathcal{B}) = 0$.

An interesting case of a nonempty $\mathcal{B}$ is that $\mathcal{B}$ is a discrete set. In such a case, we have $P(X \in \mathcal{B}) = 0$ if F is continuous. Let (z_{ij}, y_{ij}), $i \in K_1$ and $j \in K_2$, be all the possible values of (Z, Y), where K_1 and K_2 are two index sets. Then $\mathcal{B} = Q \setminus \cup_{i,j} (z_{ij}, y_{ij}]$. Notice that if $(z_{ij}, y_{ij}) = (i, i + 1 - 1/j)$, $i \geq 1$ and $j \geq 2$, then $\mathcal{B}$ is a discrete set of all positive integers. Here "$\setminus$" stands for set minus. In a follow-up study, Z stands for the age of a patient at the enrollment and Y the age at the termination of the study. Thus it is possible that in a follow-up study $\mathcal{B}$ is a nonempty discrete set with $P(X \in \mathcal{B}) = 0$, as it is reasonable to assume that the lifetime distribution is continuous.

If $P(X \in \mathcal{B}) = 0$, in general, the GMLE of F is not asymptotically normally distributed and is not asymptotically efficient (see Section 3). We propose a modified GMLE and show that it is efficient under an additional assumption. We conjecture that it is still asymptotically normally distributed and asymptotically efficient under the assumption $P(X \in \mathcal{B}) = 0$.

The organization of the current manuscript is as follows. The modified GMLE is proposed in Section 2. In Section 3 we present an example such that the GMLE is not asymptotically normally distributed and is not asymptotically efficient but the modified GMLE is. In Section 4, we make some comments.

2. Modified GMLE

Let (L, R) be the endpoints of the random interval $\mathcal{I}$ in (1.1). Let I_i, $i = 1, ..., n$, be a random sample from $\mathcal{I}$, with endpoints (L_i, R_i). We call a nonempty finite intersection B of I_i's an *innermost interval* (II) if $B \cap I_k = B$ or $\emptyset$ for all k. Let $B_1, ..., B_M$ be all the distinct

innermost intervals induced by these I_i's and assume that $x < y$ for each $x \in B_{i-1}$ and each $y \in B_i$, $i = 2, ..., M$. An innermost interval A is called a *modified innermost interval* (m-II) if it is either a singleton set or B_1 or B_M. Let $A_1, ..., A_m$ be all the distinct m-IIs and assume that $x < y$ for each $x \in A_{i-1}$ and each $y \in A_i$, $i = 2, ..., m$.

The modified GMLE (m-GMLE), $\hat{F}$, of F is defined by

$$\hat{F}(x) = \sum_{A_j \subset (-\infty, x]} \hat{s}_j,$$

where $(\hat{s}_1, ..., \hat{s}_m)$ maximizes the modified likelihood function

$$\mathcal{L}(s_1, ..., s_m) = \prod_{i=1}^{n} \sum_{j=1}^{m} s_j \mathbf{1}(A_j \subset I_i), \quad \text{where } s_j \geq 0 \text{ and } \sum_{j=1}^{m} s_j = 1. \qquad (2.1)$$

Here $\mathbf{1}(\cdot)$ is the indicator function. The m-GMLE can be derived by an iterative procedure as follows. At step 1, let $s_j^{(1)} = 1/m$ for $j = 1, ..., m$. At step h,

$$s_j^{(h)} = \sum_{i=1}^{n} \frac{1}{n} \frac{\mathbf{1}(A_j \subset I_i) s_j^{(h-1)}}{\sum_{k=1}^{m} \mathbf{1}(A_k \subset I_i) s_k^{(h-1)}}, \quad j = 1, ..., m, \text{ and } h \geq 2.$$

Stop at convergence and the limit $\lim_{h \to \infty} s_j^{(h)}$ is the m-GMLE of s_j.

Thus the m-GMLE redistributes the mass among uncensored observations, or the interval $(-\infty, R_{(1)}]$ if $(-\infty, R_{(1)}) = (L_i, R_i)$ for some $i \in \{1, ..., n\}$, or $(L_{(n)}, \infty)$ if $(L_{(n)}, +\infty) = (L_i, R_i)$ for some $i \in \{1, ..., n\}$, where

$$R_{(1)} = \min\{R_i : i = 1, ..., n\} \text{ and } L_{(n)} = \max\{L_i : i = 1, ..., n\}. \qquad (2.2)$$

Denote $t_i = \max A_i$, $i = 1, ..., m-1$. Denote $\delta_{ij} = \mathbf{1}(A_j \subset I_i)$ and $\hat{\mathbf{s}} = (\hat{s}_1, ..., \hat{s}_{m-1})^t$, where $\mathbf{s}^t$ is the transpose of $\mathbf{s}$. Let $\hat{\Lambda}$ be the $(m-1) \times (m-1)$ dimensional empirical information matrix with the (i, j)th entry

$$\sum_{h=1}^{n} \frac{1}{n} (\delta_{hi} - \delta_{hm})(\delta_{hj} - \delta_{hm}) / (\sum_{k=1}^{m} \delta_{hk} \hat{s}_k)^2.$$

4

Note that $\hat{F}(t_k) = \sum_{j=1}^k \hat{s}_j = \mathbf{e}_k^t \hat{\mathbf{s}}$, where $\mathbf{e}_k$ is a $(m-1) \times 1$ vector with the first k entries being unity and the others all zero, $k = 1, ..., m - 1$, an estimator of the variance of $\hat{F}(t_k)$ is

$$\hat{\sigma}^2_{\hat{F}(t_k)} = \mathbf{e}_k^t \hat{\Lambda}^{-1} \mathbf{e}_k / n. \qquad (2.3)$$

Recall that the GMLE $\tilde{F}$ of $F(x)$ can be obtained by $\tilde{F}(x) = \sum_{B_j \subset (-\infty, x]} \tilde{w}_j$ for all x, where $(\hat{w}_1, ..., \hat{w}_M)$ maximizes the generalized likelihood function

$$L = L(w_1, ..., w_M) = \prod_{i=1}^n [\sum_j \mathbf{1}(B_j \subset I_i) w_j], \text{ with } w_i \geq 0 \text{ and } \sum_{i=1}^M w_i = 1. \qquad (2.4)$$

Turnbull (1974) shows that the GMLE of $(w_1, ..., w_M)$ is a solution to the self-consistent equation

$$w_j = \sum_{i=1}^n \frac{1}{n} \frac{\mathbf{1}(B_j \subset I_i) w_j}{\sum_{k=1}^M \mathbf{1}(B_k \subset I_i) w_k}, \quad j = 1, ..., M, \ w_i \geq 0 \text{ and } \sum_{i=1}^M w_i = 1. \qquad (2.5)$$

A solution $(\hat{w}_1, ..., \hat{w}_M)$ to (2.5) is called an SCE of $(w_1, ..., w_M)$ and an estimator $F_1(x) = \sum_{B_j \subset (-\infty, x]} \hat{w}_j$ is called an SCE of $F(x)$ if $(\hat{w}_1, ..., \hat{w}_M)$ is an SCE of $(w_1, ..., w_M)$. Both the GMLE and the m-GMLE are SCEs. These two estimators are the same when the GMLE puts zero mass on all the innermost intervals which are not m-IIs. In general, they are different.

Under the assumption $P(X \in \mathcal{B}) = 0$ and an additional assumption that X takes on finitely many values, say $x_1, ..., x_m$, we can show that the m-GMLE is efficient. Let $A_i = \{x_i\}$ and $s_i^o = P(X = x_i)$ $i = 1, ..., m$. Then $s_i^o > 0$. Under the above assumptions, with probability one, for n large enough, the random sample contains all the A_is. In view of the likelihood function (2.1), the problem reduces to parametric estimation of a multinomial distribution function with parameter $\mathbf{s}$. The m-GMLE of $\mathbf{s}$ is the MLE of $\mathbf{s}$ in this parametric estimation problem. Since $s_i^o > 0$ for all i, by the standard large sample theory (see *e.g.*, Ferguson (1996)), the MLE of $\mathbf{s}$ is consistent, asymptotically normally distributed and asymptotically efficient. The asymptotic covariance matrix can be estimated by the sample information matrix $\hat{\Lambda}^{-1}$. This justifies the use of formula (2.3). An explicit form of the inverse of the information matrix of a self-consistent estimator is given in Turnbull (1974). Since the m-GMLE is also a self-consistent estimator, the formula is applicable to the m-GMLE.

Remark We conjecture that the m-GMLE is asymptotically normally distributed and asymptotically efficient if $P(X \in \mathcal{B}) = 0$. The above paragraph confirms it with the additional assumption that $\mathcal{O}_F$ is finite. We can further prove the conjecture under the additional assumption that $\mathcal{O}_F$ consists of isolated points or $\mathcal{B}$ is a union of mutually disjoint intervals $(u_i, v_i]$s. We decide not to present the latter proof but refer them to a technical report (Yu and Wong (1998)), as it is not as short as the above paragraph but still needs an additional assumption.

3. A Simple Example

We now give an example that the GMLE $\tilde{F}$ of $F(x)$ is not asymptotically efficient and $\sqrt{n}(\tilde{F} - F)$ does not converges in distribution to a Gaussian process, but the m-GMLE does.

Suppose that in a DC model, $P((Z, Y) \in \{(0.5, y) : y \in [2, 3)\}) = g_1$ and $P((Z, Y) \in \{(z, 8) : z \in (3, 4]\}) = g_2$, where $g_1 + g_2 = 1$; $F(x) = p_1 \mathbf{1}(x \geq 1) + p_2 \mathbf{1}(x \geq 5)$, where p_1 and $p_2 > 0$. Then (L, R) takes values $(1, 1)$, $(5, 5)$, $(-\infty, y)$ and (z, ∞), where $y \in (3, 4]$ and $z \in [2, 3)$. Given a random sample of size n from (L, R), there are N_1 $(1,1)$'s, N_2 $(5,5)$'s, N_3 intervals of form $(-\infty, y)$'s and N_4 intervals of form $(z, +\infty)$'s.

Note that in this case assumption (1.4) is violated, as $Q = [1, 5]$ (see (1.4)) but $K(3-) = P(Z < 3 \leq Y) = 0$; however, $P(X \in \mathcal{B}) = 0$ as $\mathcal{B} = \{3\}$.

We now derive the GMLE and the m-GMLE. With probability one, if n is large enough, the innermost intervals are $[1,1]$, $(y_o, z_o]$ and $[5,5]$ and the m-IIs are $[1,1]$ and $[5,5]$, where y_o is the largest L_is among all $L_i < 3$ and z_o is the smallest R_i among all $R_i > 3$. Let $U_n = \frac{N_3}{N_2 + N_3} - \frac{N_1}{N_1 + N_4}$,

$$\tilde{F}_1(x) = \frac{N_1}{N_1 + N_4} \mathbf{1}(x \geq 1) + U_n \mathbf{1}(x \geq 3) + \frac{N_2}{N_2 + N_3} \mathbf{1}(x \geq 5),$$
$$\hat{F}(x) = \frac{N_1 + N_3}{n} \mathbf{1}(x \geq 1) + \frac{N_2 + N_4}{n} \mathbf{1}(x \geq 5), \tag{3.1}$$
$$\tilde{F}(x) = \tilde{F}_1(x) \mathbf{1}(U_n \geq 0) + \hat{F}(x) \mathbf{1}(U_n < 0).$$

Verify that $\tilde{F}$ and $\hat{F}$ are the GMLE and m-GMLE of F, respectively. It follows from the strong law of large number (SLLN) that the three estimators in (3.1) are all consistent. Note

6

that $N_1 + N_3$ has a binomial distribution $Bin(n, F(2))$. Thus $\hat{F}(x)$ is asymptotically efficient for all x, and $\sqrt{n}(\hat{F} - F)$ converges in distribution to a Gaussian process.

Let $p = F(4) - F(2)$, then the m-GMLE of p is $\hat{p} = \hat{F}(4) - \hat{F}(2)$ and the GMLE of p is $\tilde{p} = \hat{F}(4) - \tilde{F}(2)$. In order to show that the GMLE $\tilde{F}$ is not efficient and $\sqrt{n}(\tilde{F} - F)$ does not converges in distribution to a Gaussian process, it suffices to show $\tilde{p}$ is not asymptotically efficient and is not asymptotically normally distributed. This is done next.

Note that $\hat{p} = 0$, thus $\text{var}(\hat{p}) = 0$. However, $\tilde{p} = U_n \mathbf{1}(U_n \geq 0)$ and $\sqrt{n}U_n$ converges in distribution to U, a normal random variable with mean 0 and standard deviation $\sigma > 0$, which can be obtained by the delta method. That is, $\text{var}(\hat{p}) < \text{var}(\tilde{p})$. Thus the GMLE $\tilde{p}$ is not asymptotic efficient. Moreover, $\sqrt{n}\tilde{U}_n \mathbf{1}(U_n \geq 0)$ converges in distribution to $U\mathbf{1}(U \geq 0)$, which is not a normal random variable.

4. Discussion

Under assumption (1.4) and some additional assumptions made in Gu and Zhang (1993), both the GMLE and the m-GMLE have the same asymptotic properties as both of them are SCEs. If $\text{P}(X \in \mathcal{B}) = 0$, then both of them are uniformly strongly consistent (see Yu and Li (1998)).

The m-GMLE has two advantages over the GMLE. Under the assumption $\text{P}(X \in \mathcal{B}) = 0$, the GMLE is not efficient but we conjecture that the m-GMLE is. In the end of Section 2, the conjecture is confirmed in the case that X takes on finitely many values but the censoring vector can be arbitrary. In application, there is a computational feasibility problem in obtaining the GMLE using the self-consistent algorithm if the sample size is large. It is then desirable to reduce the number of parameters to be estimated. The second advantage of the m-GMLE over the GMLE is that it has less parameters to estimate.

When $\text{P}(X \in \mathcal{B}) > 0$, F is not identifiable on $[0, +\infty)$. Thus both the GMLE and the m-GMLE are not consistent on $[0, +\infty)$. However, the GMLE is consistent at each observation, whereas the m-GMLE is not. Thus when the GMLE assigns to an II which is not an m-II a mass which is about the same as the mass to an m-II, it may be an indication that $\text{P}(X \in \mathcal{B}) > 0$. In such a case, it is better to use the GMLE. However, we do expect that a

7